R

Radioactivity in the Marine Environment

Jordi Vives i Batlle
Biosphere Impact Studies Unit, Belgian Nuclear Research Centre, Mol, Belgium

Article Outline

Glossary
Definition of the Subject and Its Importance
Introduction
Radionuclide Speciation in Seawater
Radionuclide Interactions In the Water Column
Radionuclides in Sediments
Behavior of Radionuclides in Estuaries
Cycling in the Foodweb
Radiological Implications of the Marine Environment
Conclusions
Future Directions
Bibliography

Glossary

Absorbed dose Quantity of energy imparted by ionizing radiation to unit mass of matter, such as tissue – Units of Gray, 1 Gy = 1 J Kg^{-1}.

Activity concentration The activity of radionuclides incorporated in a particular material per unit mass or volume – Units of Bq kg^{-1}, Bq m^{-3}.

Allometric scaling The power relationship of average body mass with the rates of many metabolic processes.

Analogue radionuclide Radionuclide which has similar environmental behavior to another.

Bioavailability Fraction of a substance that can be taken up by living organisms, depending both on chemical properties of the substance and the physiological status of the organism.

Bioindicator Biological species whose presence or absence may be characteristic of environmental conditions in a particular area of habitat.

Biokinetic studies Studies of the dynamic exchange of substances between organisms and their surrounding environment.

Biological half-life The time that it takes for a substance incorporated in an organism to lose half of its concentration by natural processes of elimination.

Bioturbation Stirring or mixing of sediment or soil by organisms, resulting in the remobilization of radionuclides.

Broken arrow An accidental event that involves the loss of nuclear weapons or nuclear components but which does not create the risk of nuclear war.

Colloid Substance dispersed as small particles in water, typically 5–200 nm in diameter. Electrostatic surface forces are the key factor in determining their properties.

Concentration factor Equilibrium parameter quantifying the capacity of an organism to concentrate a specific radioelement, expressed as the ratio of activity concentration in an organism (Bq kg^{-1}, fresh weight) to activity concentration in water (Bq m^{-3}). Units of m^{3} kg^{-1}.

Distribution coefficient (*K*$_{\mathbf{d}}$) Equilibrium parameter used to quantify the partition of a radioelement between solid (soil or sediment) and liquid phases (overlaying or interstitial water), expressed as the ratio of activity concentration in solids (Bq kg^{-1}) to activity concentration in water (Bq m^{-3}). Units of m^{3} kg^{-1}.

Dose conversion coefficient Dose rate received by an organism exposed to radiation, relative to activity concentration of the source. Units of Gy s^{-1} per Bq kg^{-1} (or Bq m^{3}).

Robert A. Meyers (ed.), *Encyclopedia of Sustainability Science and Technology*, DOI 10.1007/978-1-4419-0851-3,

Ecosystem The system formed by a biological community and its nonliving surroundings.

Foodweb A succession of organisms in an ecosystem that is linked by predator–prey relationships, transferring mass and energy from one to another.

Global fallout Residual radiation hazard from the atmospheric nuclear explosions of the 1940s to 1960s, with slowly declining global presence in the planet up to the present day.

NORM Radionuclides from naturally occurring radioactive materials released to the environment by industrial processing, e.g., mineral extraction and processing, oil and gas, phosphogypsum by-products, and fertilizers.

OSPAR The Convention for the Protection of the Marine Environment of the Northeast Atlantic, a mechanism by which 15 Governments of the western coasts and catchments of Europe, together with the European Community, cooperate to protect the marine environment of the Northeast Atlantic.

Radiation weighting factor Factor representing the relative effectiveness of different radiation types relative to X- or gamma-rays, in producing biological effects of significance.

Remobilization The return of a radionuclide to circulation within an ecosystem, such as return to the water column of radionuclides which had been locked in sediments.

Scavenging Removal of substances from the water column by the action of settling particles and/or organic matter.

Sediment mixing Mixing of sediment and water at the interface between the two, by processes such as bioturbation and water turbulence.

Speciation Partitioning of an element amongst defined species in a chemical system.

Definition of the Subject and Its Importance

The sea is a complex system containing different components (water column, suspended particulates, colloids, sediments, and organic matter) and inhabited by life forms at multiple scales (from plankton to large mammals), undergoing complex interactions.

With the arrival of nuclear technology in the late 1940s, a variety of man-made radionuclides have entered the marine environment, either as a result of military operations, industrial discharges, medical releases, or nuclear accidents. This has resulted in their widespread distribution, cycling across the sea and uptake by biota, both locally (in the vicinity of discharge points) and globally.

Studies of the different ways radioactive substances interact with the marine environment fall into the domain of marine radioecology. In the 1950s and 1960s, this discipline focused on studying how the global fallout from nuclear weapons testing and satellite burn-up entered the oceans and was distributed within the water column, reaching the seabed sediments and the food chain. Radionuclides of interest were the relatively long-lived plutonium (Pu), americium (Am), radiocesium (Cs), and radiostrontium (Sr), the fission products from nuclear detonations.

From the late 1970s to 1980s, more studies explored the interaction mechanisms of radionuclides with sediments, suspended particulates, and marine colloids. A key focus for marine radioecology research at that time was the input from European reprocessing plants to the seas of the northern hemisphere. Most significantly, the marine discharges from the reprocessing plants at Sellafield, UK, and La Hague, France resulted in enhanced radionuclide levels in the Irish Sea and English Channel/North Sea respectively. Fundamental to investigations at this time was to understand the chemical speciation of the transuranic radionuclides, predominantly plutonium and americium, and the mechanisms of how they concentrated in sedimentary deposits and how they became remobilized.

A key driver for this phase was the estimation of radiation doses to humans arising from the consumption of contaminated marine foodstuffs. However, beyond the purely radiological protection point of view, substantial understanding was sought and gained on the potential of radionuclides as tracers in the study of marine and coastal processes, as well as understanding the transfer of radionuclides to the local communities of animals and plants.

The Chernobyl accident in 1986 resulted in a global release of radionuclides. Although in the Irish and North Seas the main influence continued to be the input from reprocessing plants, the Baltic and Black Seas became affected by the Chernobyl accident, with ^{90}Sr, ^{134}Cs, ^{137}Cs, and ^{239,240}Pu entering these environments. Meanwhile, during the 1990s, Irish Sea studies shifted focus to the ^{99}Tc discharges arising from the Enhanced Actinide Removal Plant (EARP) in Sellafield, with ^{99}Tc discharges reaching their peak between1994 and 1996. As a result, levels of ^{99}Tc in organisms such as lobsters found off the coast of Sellafield reached their highest levels in the late 1990s and early 2000s. During this time, transuranium discharges decreased several orders of magnitude below their peak discharge rates of the mid-1970s.

Before the Fukushima accident in Japan in 2011, there were no more major releases of radioactivity into the marine environment. In a period marked by concerns on long-term environmental sustainability came the realization that radionuclides persist in the marine environment for long periods of time. Research addressed in more detail the interaction of radionuclides with individual organisms and foodwebs, striving for a global understanding at a species and ecosystem level. The traditionally anthropocentric view of radiological protection was replaced by a more ecocentric approach. There was a realization that, even if humans are protected, the environment is not necessarily protected because marine organisms inhabit areas where humans cannot reach. This called for the development of an international system of radiological protection for the environment, and new studies of the transfer of radionuclides to marine biota.

Recently, dynamic situations such as accidental scenarios, decommissioning discharges and NORM releases from offshore oil operations have attracted scientific interest. These situations are characterized by irregular discharge patterns, requiring dynamic modeling of the transfer of radionuclides to marine organisms. There is also interest in low-level ionizing radiation effects to biota, and it is felt that Radioecology needs to be applied in a world of multiple contaminants, exerting their combined stress on interdependent species and entire ecosystems. A number of collaborative international projects are developing in this direction.

Radioactivity in the marine environment is here to stay, and there is an onus to gain a scientific understanding of its implications, as part of a drive to preserve the quality of this environment for future generations.

Introduction

Artificial radioactivity reaches the world's oceans from various sources: releases from nuclear power plants and reprocessing facilities connected to coastal areas directly or via waterways, the worldwide fallout from nuclear weapons tests, dumping of waste to the seafloor or accidental situations such as satellite burn-up, "broken arrow" situations, or the sinking of nuclear submarines.

The historical sources of radioactivity into the environment, detailing natural radioactivity, global fallout, accidents, and NORM and discharges from the nuclear industry to the marine environment have been abundantly described elsewhere [1–8], so they will only be summarized briefly.

Historically, one of the most significant inputs with relevance to the marine environment is the worldwide fallout of nuclear weapons tests, resulting in the global inventory of man-made radioactivity in the world's oceans [9, 10], to which one must add accidents involving space satellites [11, 12] and lost nuclear weapons [2, 13–16]. The earliest weapons tests provided in fact the first opportunities to investigate at close quarters the impact of man-made radionuclides in the marine ecosystem [17–19].

Nuclear accidents represent another source of radioactivity to the marine environment. The Chernobyl accident affected indirectly several marine areas [20–23]. The two major nuclear accidents in 1957, Kyshtym and Windscale, also affected the marine environment to some degree, but they are not believed to have been a major source of contamination to the marine environment [6]. At the time of writing this article (April 2011), it is too early to evaluate the impact of the 2011 nuclear accident at the Fukushima plant (Japan), though it is clear that this event resulted in the direct release of radionuclides to sea, chiefly ^{131}I and ^{137}Cs.

Routine operations from the nuclear industry have been a source of radioactive contamination to sea since the early 1950s. The most important contribution is from the reprocessing of spent nuclear fuel in purpose-built facilities discharging waste to sea. The importance of the Sellafield site discharges to the Irish Sea has resulted in numerous radioecological studies being performed, examples of which are described here [24–28]. Another significant reprocessing site which has been extensively studied from the marine discharges point of view is the La Hague reprocessing plant in France [29, 30]. In the Arctic Ocean, run-off from contaminated sediment in the Ob and Yenisey river system connected to the Mayak, Krasnoyarsk, and Tomsk nuclear facilities and the Semipalatinsk test site has also become a well-studied source of radionuclides [31].

Lastly, historical dumping of radioactive waste [32] has been a controversial source of radioactivity release into the world's oceans. Between 1946 and 1982, packaged low-level radioactive waste (LLW) was dumped at more than 50 sites in the northern part of the Atlantic and Pacific Oceans. Beta-gamma emitters represented the majority of the total radioactivity in that waste, but alpha-emitters such as plutonium and americium were also disposed of by this method [33].

By the above means, an extensive list of man-made radionuclides has been introduced to the marine environment in varying amounts. This article cannot cover every one of them. Due to their availability in the environment, their radiological importance, and the amount of environmental data available, the four radionuclides to be considered further are ^{99}Tc, ^{137}Cs, Pu (as the α-emitters ^{238}Pu and $^{239+240}$Pu, as well as the β-emitter ^{241}Pu), and ^{241}Am. These radionuclides can be grouped broadly into two categories: those behaving as dissolved in the water column (Tc and Cs) and those with a strong affinity for suspended particulates and sediments (Pu and Am).

In sea water, almost all of the naturally occurring elements are present in variable quantities. The concentrations of the major constituents are found to be directly proportional to the salinity of the water. The more soluble Tc and Cs radionuclides follow the same behavior, and so they are termed "conservative." The sediment-seeking Pu and Am are classed as "nonconservative" radionuclides. The residence time within the seabed for these non-conservative radioelements will generally be longer than the hydrological transit time following release or remobilization from the sediments.

After radionuclides have entered the water, physicochemical changes will occur, followed by dilution and dispersion further afield by the action of oceanographic processes. However, it is also possible for radionuclides to become accumulated in certain parts of the marine environment, such as sediments, through the processes of scavenging and particle deposition, as well as becoming biologically concentrated. A conceptual representation of the key processes is given in Fig. 1.

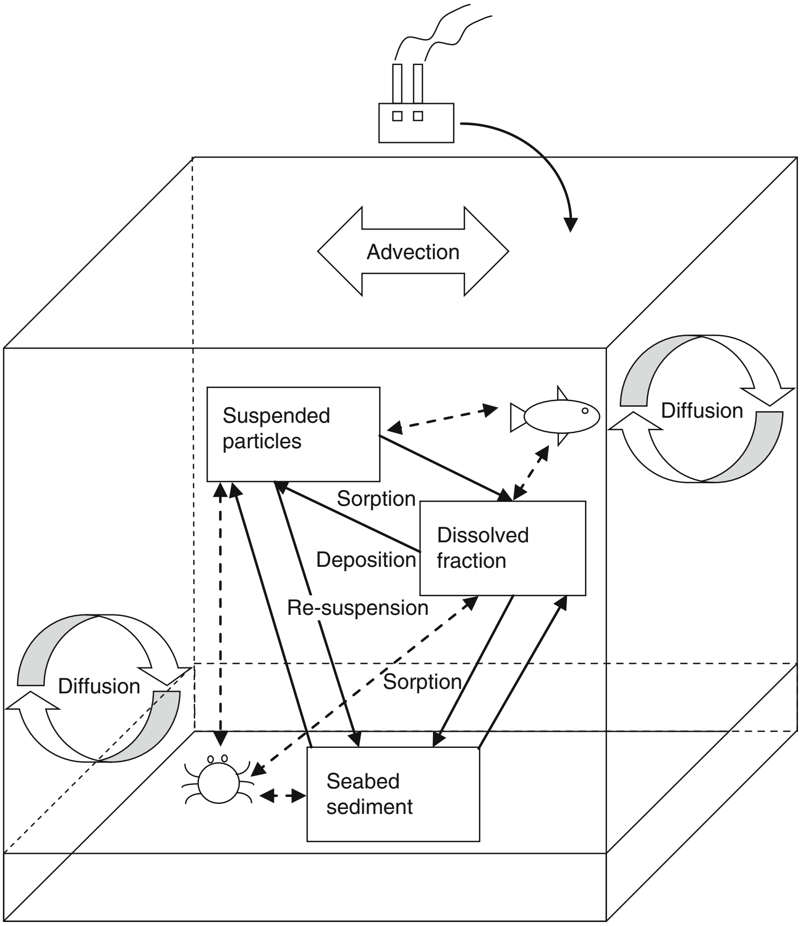
Whether radionuclides from a specific source are associated with particles, colloids, or various chemical species in the water depends on the type of source and conditions prevailing in a particular environment [7, 34]. For example, plutonium from global fallout is mainly associated with submicron iron oxide particles; in fallout from surface tests in the Marshall Islands, plutonium was mainly attached to calcium hydroxide particles [35]; Pu in effluents from nuclear reprocessing plants are associated with particulate material or colloids [36].

Ultimately, the transport, distribution, and biological uptake of radionuclides in the marine environment depends heavily on the physicochemical forms of radionuclides (i.e., their speciation) in the discharged effluents, as well as on the transformation processes that occur after entering coastal waters. For this reason, this article will cover in some detail radionuclide speciation, a "hot topic" in the aquatic environment where trace metals are concerned.

Trace element speciation in the marine environment is different from speciation under freshwater conditions. This is due to varying features such as salinity, pH, dissolved organic carbon and the size of particulates in the water column, ranging from nanocolloids to micron-sized particulates. Many reviews on radionuclide speciation concentrate on particle size considerations, i.e., physical speciation. In this entry, consideration will be given to the actual chemical speciation, which refers to the molecular form of the radionuclides in the marine environment.

Radionuclide Speciation in Seawater

Conservative radio elements, like ^{90}Sr, ^{137}Cs, and ^{99}Tc, have a relatively low (but measurable) capacity for adsorption onto solid particles. They readily disperse upon release to the marine environment, and therefore can be used as tracers for oceanographic processes.



Radioactivity in the Marine Environment. Figure 1
Conceptual representation of the processes occurring when radionuclides enter the marine environment

On the other hand, transuranium elements, like plutonium and americium, are not efficiently dispersed away from the source point but, instead, are more efficiently carried onto the seabed by scavenging particles, becoming incorporated onto sediment for protracted periods of time. ^{241}Am is a special case as additional quantities will be formed by the decay of ^{241}Pu if this radionuclide is present.

The above picture is somewhat simplified, and becomes more complicated by the fact that certain radionuclides, such as plutonium, can have different chemical forms, or species, capable of coexisting in varying proportions in the water column. Some of these species actually have low particle reactivity, behaving conservatively, whereas others can have a high affinity for sediment and behave nonconservatively. In this section, the speciation of Pu and ^{241}Am is compared on one hand, and ^{99}Tc and ^{137}Cs on the other, to illustrate how this phenomenon affects their environmental behavior.

Plutonium and Americium Speciation

It is well known that, in aqueous solution, plutonium can exist in the oxidation states +3, +4, +5, and +6 [37]. In the absence of complexing agents, these oxidation states form the chemical species Pu^{3+} and Pu^{4+} (the “reduced,” or Pu(III,IV) group) and PuO_2^{+} and PuO_2^{2+} (the “oxidized,” or Pu(V,VI) group), respectively. These two oxidation state groups of plutonium exhibit very different sediment sorption properties, with reduced plutonium possessing an affinity for sediment approximately two orders of magnitude higher than oxidized plutonium, relative to concentration in seawater.

Such species can be interconverted from one to another by means of oxidation–reduction reactions, coexisting in equilibrium in the aquatic environment [9, 38–41]. Theoretical calculations show that in a pure aqueous solution (without suspended particles or colloidal matter) PuO_2^+ is, by far, the predominant form [42–45]. The insolubility of $Pu(OH)_4$ is the limiting factor of the net solubility of plutonium in oxic natural waters, making $Pu(V)O_2^+$ the most stable oxidation state [46].

Americium in aqueous solution can also exist in the oxidation states +3, +4, +5, and +6 but the oxidation state +4 is stable only in the presence of concentrated H_3PO_4, $K_4P_2O_7$ and fluoride solutions [47]. In the absence of complexing agents, the oxidation states +3, +5, and +6 can occur in the form of the hydrated ions Am^{3+}, AmO_2^+, and AmO_2^{2+}. However, AmO_2^+ and AmO_2^{2+} are quickly reduced to Am^{3+}. Hence, the chemical speciation of americium, as well as its environmental behavior, is less complex than that of plutonium and Am^{3+}, likely to be in the form of a hydroxide, is the predominant species of americium in seawater [48].

Understanding the environmental behavior of plutonium in seawater requires consideration of the complex interactions of Pu (III,IV) and Pu (V,VI) within a heterogeneous water column where two distinct solid phases (particulate and colloidal) coexist, as well as interactions with seabed sediments. The situation is fully described elsewhere [49, 50]. Environmental factors such as the physical nature of the discharge (i.e., particulate vs. soluble radionuclides in the original effluent), salinity [51], the presence of dissolved organic carbon [52], suspended particulate load and colloids [53] can alter the balance, increasing the proportion of the "reduced" species in the water column [50, 54, 55].

The above can be illustrated with the following example. If plutonium is released in an insoluble, particulate form in a relatively shallow coastal area, it will likely be incorporated in the seabed sediments, with the relatively smaller oxidized fraction traveling farther afield. The result is that the percentage of oxidized plutonium in seawater will increase over the first few kilometers away from the outlet, in the direction of the dispersion. This is the case of the Sellafield reprocessing discharges where plutonium releases occurred in particulate form and the bulk of the historical inventory is incorporated into sedimentary deposits [48, 50]. However, if plutonium is released in oxidized form, this oxidized fraction will dominate throughout the dispersion pathway. For americium, the situation is much simpler, since the oxidized form Am(VII) is unstable and most of the Am in seawater is reduced. Hence, this radionuclide will largely be incorporated into particles and sedimentary deposits.

It therefore follows that most of the plutonium in open waters (well removed from source terms) is to be found in the oxidized form Pu(V,VI), behaving like a conservative tracer which can be transported within the dissolved phase by advection-dispersion processes. This has been verified experimentally in the open waters of the Irish Sea [37, 50, 56, 57], semi-enclosed seas and shallow continental shelf waters such as the Mediterranean Sea [41, 50], the benthic environments of Bikini and Eniwetok Atolls [9, 58], the northwest Greenland continental shelf [55], and the English Channel [59]. In contrast, ^{137}Cs and the similarly behaved ^{99}Tc are readily dispersed, with relatively little quantity residing in sediments compared with plutonium and americium [60–62].

A smaller fraction of the plutonium in open waters is in a reduced form, and this is partly associated with suspended particulate matter. This scenario becomes somewhat complicated by the fact that Pu(IV) in seawater can reoxidize to Pu (V, VI) simultaneously with the slower opposite reaction, with rates of the order of 10^{-1} and 10^{-2} days^{-1}, respectively [49, 59]. In contrast, all of the americium in open waters is in a reduced form, characterized by a strong affinity for suspended particulate matter, colloids and sedimentary deposits.

It follows from the above that, although it is true that they have an affinity for sediments [63–66], the fate and transport of plutonium and americium depend heavily on oceanographic considerations as well as the nature of the discharges themselves. In addition, intermixing of the different fractions of Pu (oxidized and reduced) takes place over long timescales.

Technetium and Radiocesium Speciation

The speciation of the conservative elements ^{99}Tc and ^{137}Cs is relatively simple compared with Pu and Am. The effect of technetium chemistry on its environmental behavior has been the subject of numerous studies and reviews [67–76], from which the following summary may be derived.

The pertechnetate anion TcO_4^-(VII), a soluble species, is the form most likely to be discharged in low active liquid effluents from nuclear fuel reprocessing facilities and, thus, the dominant form of technetium in the marine environment [77]. Strong reducing conditions are required to overcome the inherent stability of TcO_4^-, forming the particle reactive species technetium dioxide TcO_2(IV). For instance, these conditions can be found in oxygen-depleted sediments, but the difficulty in reducing Tc(VII) means that reduced Tc(IV) may only be found where pH and redox conditions permit it. Once reduced, technetium may exist as insoluble technetium dioxide or technetium sulfide species. Environmental Tc(IV) can also be stabilized by organic matter. Precipitation or association with humic acids may occur in sediments.

Given its chemical behavior, the dominant "oxidized" species of technetium, TcO_4^-, behaves as a soluble compound. As such, it has a high mobility throughout the environment. In contrast, "reduced" Tc(IV) behaves as an insoluble compound and is much less mobile.

^{137}Cs is a Group I metal existing as the monovalent cation Cs^+, its only environmentally available oxidation state. Speciation studies carried out in the Baltic Sea confirm that water-soluble cesium constitutes most (60–99%) of the total ^{137}Cs in seawater, predominantly as Cs^+, both in near–shore and open waters. The remainder of radiocesium in the water column is in particulate form, varying between 1 and 30% depending on environmental conditions [78]. The uptake of Cs onto suspended matter and sediment is proportional to the illite content of the sediments, probably involving an ion-exchange reaction with potassium [79]. Hence ^{137}Cs is stably bound by crystalline minerals, more so than by organic matter [80].

For the above reasons, technetium and radiocesium are considered to be long-distance tracers in the marine environment, more suitable for following hydrodynamic transport processes than their actinide counterparts. In contrast to Pu and Am, they have the ability to disperse readily in seawater [60–62]. In the Irish Sea, e.g., these radionuclides are advected, principally in a northerly direction, associated with the mean flow of water currents, leaving the area via the North Channel with a mean transit time of about 1 year [81].

Iodine is of renewed interest due to the recent Fukushima accident. This radionuclide is generally soluble in seawater, and poorly adsorbed onto particulates and sediment. Both iodide (I^-) or iodate (IO_3^-) anions can be present, but the latter predominates [82]. In this way, IO_3^-can be broadly classed as analogous to Cs and Tc in the marine environment.

Radionuclide Interactions In the Water Column

When radionuclides are released into seawater, physicochemical processes such as polymerization, disproportionation and the formation of precipitates can occur. As a result, there takes place a redistribution between the amount of radionuclide in solution and the insoluble forms partially attaching to solid surfaces. At trace concentrations, the dominant process for this attachment is primary adsorption, where a monomolecular layer of the adsorbed substance is formed on the surface of a particle. Simultaneously, radionuclides are dispersed by turbulent mixing and advective currents, undergoing significant dilution in the process.

Interactions with Suspended Particulates

Radionuclides initially released in an ionic form may undergo molecular dispersion as simple compounds. However, after some time, low solubility forms of these simple compounds can become adsorbed onto the surfaces of suspended particulate material, e.g. the fine suspended matter present in estuaries.

Varying proportions of Pu and Am are associated with the particulate phase [50, 83–85]. In the surface waters of the North Pacific and the Mediterranean, the proportion of particulate to total Pu is less than 10%, whereas in the deep waters of the Pacific it is less than 2% [86]. In coastal and estuarine environments, this proportion can be much higher [87]. This contrasts with ^{90}Sr, ^{137}Cs, and ^{99}Tc, which have lower fractions in suspension.

Since the soluble and particulate forms of these radionuclides behave quite differently, it is important to know the partitioning between the aqueous phase and the solid phase (suspended particulates or sediment) on a radionuclide by radionuclide basis. For simplicity, this is represented by an equilibrium ratio, the distribution coefficient or K_d, which include implicitly the complex processes involved:

$$K_d(\mathrm{L\ kg^{-1}}) = \frac{\text{Activity concentration per unit mass of solid (Bq/kg dry weight)}}{\text{Activity concentration per unit volume of water (Bq/L)}}$$

K_d values for the various elements in seawater are provided in the literature, duly compiled by the International Atomic Energy Agency (IAEA) on the basis of best estimates and empirical data for various environments [88]. A summary of representative K_ds for Cs, Tc, Am, and Pu in open oceans and coastal environments is given in Table 1.

As explained previously, plutonium has different chemical forms. Pu(III,IV) has a K_d one order of magnitude higher than given in Table 1, of the order of 10^3 m^3 kg^{-1}, similar to the K_d for americium [37, 45, 50]. In contrast, the IAEA recommends a K_d for technetium of 10^{-1} m^3 kg^{-1} for both pelagic clays and coastal sediments [88, 89], derived from Irish Sea measurements [90]. This K_d may be taken to represent the most oxygenated marine sediments, but higher values should be applied to anoxic sediments with high organic matter content [67]. For radiocesium, measured K_ds are in the range of 10^0–10^1 m^3 kg^{-1} (IAEA recommended value = 3.5×10^0 m^3 kg^{-1}) [88, 89].

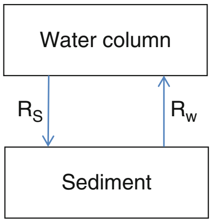
The K_d is a somewhat crude simplification of reality, since time is required for complete equilibration to be attained between the aqueous and solid phases. Typically, for suspended particulate concentrations in the order of 10^{-3}–10^{-2}g L^{-1}, adsorption equilibrium is achieved within seconds to minutes, but processes other than sorption may considerably delay the adjustment between the soluble and the suspended load phases. In such cases, a so-called effective K_d can be derived, which should not be interpreted simply as an equilibrium adsorption parameter [45].

When effluent released in particulate form travels away from its source, the K_ds for plutonium and americium will decrease with distance from the outfall. In the Irish Sea, an order of magnitude exponential fall has been observed in transit between the eastern and the western regions, as the partitioning between liquid and solid fractions re-equilibrates [45, 50]. Sediment resuspension and dissolution continue to influence the dispersion of Pu along the route [91].

For seabed sediments, equilibration takes much longer to be achieved. Some understanding of the relevant mechanisms can be gained by a simple model involving two compartments: water column and sediment, with uptake and release rates R_s and R_w mediating between them (Fig. 2). The processes to consider are: depletion of radionuclides adsorbed onto suspended particulates (particle scavenging), molecular diffusion, pore water mixing and bioturbation, i.e., the disturbance of sediment by living organisms, modeled effectively as a diffusive process.

Radioactivity in the Marine Environment. Table 1
Representative K_d values for Cs, Tc, Am, and Pu in open oceans and coastal environments [88]

	K_d value (m^3 kg^{-1})	
Radionuclide	Open waters	Coasts and estuaries
Cs	2×10^0	4×10^0
Tc	1×10^{-1}	1×10^{-1}
Pu	1×10^2	1×10^2
Am	2×10^3	2×10^3



Radioactivity in the Marine Environment. Figure 2
Simplified representation of the exchange of radionuclides between surface sediment and the water column [92, 93]

Previous investigators [92, 93] have given the basic equations describing the combined transfer rates for these processes:

$$\begin{aligned} R_{\mathrm{s}} &= \text{Particle scavenging} \\ &\quad + \text{Molecular diffusion} + \text{Pore water mixing} \\ &\quad + \text{Particle mixing} \\ &= \frac{1}{d_{\mathrm{w}}(1 + K_{\mathrm{d}}\alpha)} \\ &\quad \times \left(SK_{\mathrm{d}} + \frac{D + h^2 \varepsilon R_{\mathrm{T}} + R_{\mathrm{W}} K_{\mathrm{d}} \rho h (1 - \varepsilon)}{h} \right) \\ R_{\mathrm{w}} &= \text{Molecular diffusion} + \text{Pore water mixing} \\ &\quad + \text{Particle mixing} \\ &= \frac{1}{h\varepsilon \left(1 + K_{\mathrm{d}} \rho \frac{(1-\varepsilon)}{\varepsilon} \right)} \\ &\quad \times \left(\frac{D + h^2 \varepsilon R_{\mathrm{T}} + R_{\mathrm{W}} K_{\mathrm{d}} \rho h (1 - \varepsilon)}{h} \right) \end{aligned}$$

where S is the sedimentation rate, d_{w} the water layer depth, α the suspended sediment load, D the diffusion coefficient, h the thickness of the sediment layer, R_{T} the pore water turnover rate, R_{W} the sediment reworking rate, ε the porosity, and ρ the density. In the case where there is no suspended load (and hence, no sedimentation) and the sediment is assumed to be a simple solid surface (zero porosity), the ratio between the uptake and turnover rates equals the quotient between the activity, in Bq, in the sediment relative to the water:

$$\frac{R_{\mathrm{s}}}{R_{\mathrm{w}}} = \frac{\rho h}{d_{\mathrm{w}}} K_{\mathrm{d}}$$

In reality, the adsorption behavior of radionuclides is influenced by a number of environmental factors including salinity, temperature, depth, size distribution and composition of the suspended load, biological productivity, trace metal concentrations, and the presence of colloids and organic pollutants.

Interaction with Colloids

Another associative mechanism for particle-seeking nuclides initially in ionic form is the attachment to colloidal and pseudo-colloidal matter. Colloids are nanoparticles and macromolecules in the size range 1–10 kDa (or 5–200 nm) [50, 87], held apart by the mutual repulsion of their negative charge [94]. Colloids are in fact abundant in both shelf and open ocean surface waters, with a total surface area of about 8 m^2/m^3 of sea water [95–97]. Given the high surface per unit mass of colloidal systems and the electric charge available, it is inevitable that they will be efficient in fixing radionuclides that have low solubility per se. This means that particle-reactive radionuclides released in an "insoluble" chemical form (such as Pu(III,IV) or americium) could potentially behave as if they were soluble, dispersing away with the water.

Colloidal and pseudo-colloidal particles can be inorganic (polyhydroxy complexes or polysilicates) [98] or organic, i.e., colloidal organic carbon (COC), which is composed of combined amino acids, carbohydrates, and lipids. There are even higher molecular weight components such as humic acids and their compounds, which have a high capacity to strongly and specifically complex metals. These components are formed upon the degradation of organic matter within the water column and the seabed. The resulting products are stable over protracted periods of time and exist in a wide range of sizes: in sea water, dissolved organic carbon (DOC), comprises structures with molecular weights varying from $<10^3$ to over 2×10^5.

It is known that part of this DOC exists in a colloidal form [45]. Typical concentrations of colloidal organic carbon in unpolluted ocean surface waters are about 1–5 $\times$ 10^{-3} kg of organic matter per m^3 [99, 100], similar to typical suspended particulate matter concentrations. In these environments, the percentage of plutonium which appears to be bound to COC seldom exceeds 20% [45].

In coastal regions and estuaries, colloidal particles are likely to be present at higher concentrations, explaining perhaps the increased percentage of reduced plutonium in apparent solution in these environments. Conversely, in the open waters of temperate environments, the percentage of colloidal plutonium seldom exceeds 15% [50]. In the subsurface waters at Thule (northwest Greenland) there appears to be little, if any, plutonium associated with colloidal species [55].

The effects of DOC on the sorption properties of plutonium in natural water systems have been studied in some detail [52, 101], and it has been demonstrated that DOC plays an important role in regulating the binding of plutonium (and presumably americium) to particulate matter. Furthermore, it has been shown that the K_d for reduced plutonium is strongly correlated with DOC according to an equation that is almost constant for very low COC concentrations, quasi-linear at intermediate levels and parabolic at high concentrations:

$$\frac{1}{K_d} = \frac{1}{K_d^0} + \frac{K_1}{K_s}[\mathrm{COC}] + \frac{K_2}{K_s}[\mathrm{COC}]^2$$

where $K_d{}^0$ is the limiting distribution coefficient, that is to say, the K_d value for [COC] = 0. This behavior has been reproduced in several laboratory experiments [102].

In coastal waters with high levels of colloidal and dissolved organic carbon (COC/DOC), Pu(III,IV) and americium may have a significant fraction in colloidal form, as a result of the repartitioning of radionuclides between the dissolved and the particulate phases due to competition for adsorption sites with COC/DOC. However, for some radionuclides exhibiting complex speciation, the proportion attached to particulate matter can actually increase. This has been observed experimentally for plutonium and it is due to the reduction of plutonium(V) to plutonium(IV) in such environments [50, 103].

In open waters, virtually all of the plutonium, including the fraction in a reduced chemical form, is present as a fully dissolved species [104]. This implies that long-range transport of radionuclides via a colloid pathway is unlikely, but the possibility of radionuclide interaction with some colloidal matter like naturally occurring complexing ligands, actually contributing to the overall transport of radionuclides, cannot be discounted [58].

Contrasting with the potential of ^{239,240}Pu(III,IV) and ^{241}Am to be in a colloidal state, laboratory experiments on dilution of radioactive effluent into sea water indicate that ^{90}Sr, ^{137}Cs, and ^{239,240}Pu(V) do not have such an ability to persist in colloidal form [36]. This is due to the lower chemical reactivity of these radionuclides, hence mechanisms for their binding to colloids are passive, and consequently weaker [105, 106].

Colloids are not always inherently stable and will tend to form aggregates, particularly at high salinity, resulting in coagulation of the material and the eventual deposition of radionuclides to the seabed [107]. This process appears to be biphasic, i.e., represented by the sum of two exponential functions, with an initial fast phase (0–5 h) transitioning to a slower process (5–120 h), and rate constants of 2.1×10^{-2} and 3.5×10^{-3} h^{-1}, respectively [49]. Since only ~25% of the colloidal phase coagulates via the initial fast phase, a coagulation half-time of 8.25 days (corresponding to the rate constant of 3.5×10^{-3} h^{-1}) and a "colloidal" K_d of 10^4 m^3 kg^{-1}can be used to represent the overall process at long timescales [108].

Scavenging from the Water Column

As suspended particulates and coagulated colloids sink to the bottom by the action of gravity, the radionuclides attached to them are gradually removed from surface waters. Hence, fallout plutonium and americium concentrations in surface waters have been decreasing over the decades, with preferential removal of americium with respect to plutonium.

In the Mediterranean Sea, e.g., mean residence times for plutonium and americium in surface waters are about 15 and 3 years, respectively, so only a small fraction of the fallout is still in the water column [109]. For conservative radionuclides such as Sr, Tc, and Cs the process is much slower, but after time a significant fraction of those radionuclides has also deposited onto the seabed. To give an idea of the difference between conservative and nonconservative radionuclides, average residence times of ^{90}Sr (^{137}Cs) and ^{239,240}Pu in surface waters of the world are calculated to be 27 ± 2 and 13 ± 1 years, respectively [6].

Radionuclide depletion from the water column occurs by direct precipitation or by absorption on suspended particles and their aggregates, which sink to the bottom at rates of the order of 70–400 m year^{-1} [110]. Another important mechanism for the removal of radioelements from surface water is a biological one: scavenging by biological aggregates such as marine snow and plankton fecal pellets. Studies have shown that such fecal pellets are relatively rich in plutonium and americium. Metal binding on copepod fecal pellets and marine snow appears to be generally greater than on euphausiid fecal pellets [111].

The process of decomposition and release of trace elements from zooplankton debris has been investigated in some detail [112]. Release rates of metals/ metalloids (^{65}Zn, ^{75}Se, and ^{241}Am) in the context of carbon release appear to decrease exponentially over time, with the most pronounced decreases occurring within the first 6 days. The retention half-time of ^{241}Am in fecal pellets is in the order of 50 days, compared with 1–10 days for the other elements in this study. This result helps to explain the enrichment of scavenged transuranics in fecal pellets and their short residence times in surface waters.

Overall, the mechanism of fecal pellet sinking and degradation is an important factor in the vertical transport of many radionuclides from surface to bottom waters [111, 113–115]. In fact, zooplankton grazing in the surface layers and concentration of particle-reactive radionuclides into large and dense (hence rapidly sinking) fecal pellets is the plausible mechanism whereby Chernobyl-derived radioactivity (particularly the rare earths) entering the Mediterranean was rapidly transported to 200 m in the few days after the accident [116]. This has potential implications for the recent Fukushima nuclear accident in Japan, which discharged during spring time into the Pacific coast off the island of Honshu, a moderately high productivity ecosystem of complex oceanography [117–119]. In this area, radionuclides have the potential to be entrapped in the gyres formed by the interplay between the Kuroshio and Oyashio currents, which hypothetically limit the dispersion of radionuclides from the area while driving them partially to the seabed, with increased exposure to benthic organisms.

In shelf seas, relatively high inventories of plutonium can accumulate in the continental shelf sediments, attributed to high-density particle scavenging, and contrasting with deep areas where particulate inputs from the shelf are generally low [120]. In the open oceans, particle scavenging by falling biogenic particles can generate a distinct concentration front (subsurface maximum) at depths of 100–1,000 m for plutonium and americium [121], as observed in the Northeast Pacific [122, 123], North Atlantic [124, 125], western Mediterranean [50, 126], Norwegian and Greenland Seas [127] and the central Arctic Ocean [128, 129]. The radionuclide profile in the water column generally exhibits a typical pattern with a surface minimum, a mid-depth maximum and gradual decrease with increasing depth. As time passes, this subsurface maximum tends to sink. For example, in the central Northwest Pacific Ocean, the fallout ^{239,240}Pu subsurface maximum has moved from about 500 m in the early 1970s down to 800 m by the late 1990s, as a result of these processes [130].

Following deposition, radionuclides can migrate through the sediment column, either by diffusion through pore water or its expulsion from compacted sediments, groundwater discharges through the seabed, turbulent mixing in shallow areas or the action of burrowing organisms (bioturbation). In addition, sediment-dwelling species can play a significant role in the remobilization process, as they may redistribute radionuclides from sediments by mixing and agitation [131]. In estuaries where abundant sedimentation occurs it is this, rather than diffusion, that becomes the dominant process [132].

Radionuclides in Sediments

The high Pu and Am K_d's in Table 1 exemplify how strong the affinity of these radionuclides for sediments is. Particle size is a major influence, with a greater tendency for adsorption onto the smaller particles found in mud and silts [133]. To a lesser extent, the same affinity for fine particles applies to Tc and radiocesium [73, 134–136].

Despite its predominantly conservative behavior, a certain proportion of the radiocesium also absorbs onto sediments ($K_d = 2\text{–}4 \times 10^0$ m^3 kg^{-1}, see Table 1). For this reason, a radiocesium inventory still resides in areas such as the Irish Sea, along with the transuranium radionuclides. Anionic technetium adsorbs poorly onto particulate matter, which has almost exclusively cationic sorption sites [137]. However, sediments with a high organic content can also retain technetium, especially under reducing conditions. There is a possible link with the presence of bacteria [138], as it has been suggested that the organic polymers that coat the cells of microorganisms may be responsible for technetium sorption [139].

Technetium VII may be reduced to technetium IV forming insoluble compounds such as $TcO_2.2H_2O$, coprecipitated with metallic sulfides or precipitated as Tc_2S_7 [140, 141]. The above studies show that the technetium is most likely reduced at quite a fast rate to insoluble and relatively immobile technetium dioxide by biogenic iron in sediments [142]. It is, therefore, postulated that technetium remains relatively immobile in sediments beneath the depth at which reduction occurs [68].

Radionuclide Inventories in Sediments

Of the many instances of radionuclide bioaccumulation in sediments, the Irish Sea is noteworthy due to the amount of studies published, which serve to illustrate the environmental processes at play. Over the past 4 decades, the reprocessing discharges from Sellafield have peaked and are now in decline [25, 29, 81, 84]. The sediment inventory from historic discharges has, therefore, become the dominant source to the water column [66, 136, 143–145]. In particular, the mud patch located in the immediate vicinity of the Sellafield outfall remains significant as a repository of plutonium, americium, and radiocesium from the site.

At least 85% of the sediment-bound plutonium and americium in the Irish Sea lies within this coastal strip, approximately 30-km wide, near the Cumbrian coast, buried at mid-depth (~50-cm), and only a small fraction (<7%) of the plutonium discharged annually from Sellafield leaves the area [49, 50, 65]. Accumulation in such a small zone can be explained by local water movements and the fact that most of the plutonium and americium were originally associated with particulate matter in the effluent itself [48].

The distribution of plutonium, americium, and radiocesium in the sub-tidal sediments and interstitial waters of the northeastern Irish Sea has been assessed through major surveys in the late 1970s to mid-1990s [144, 146]. A budget of Pu(alpha) based on these published observations totals 460–540 TBq, or 64–75% of the reported discharge still residing in the zone of fine sediment. The solid phase profiles of plutonium and americium have major subsurface peaks that appear to reflect the history of discharges from Sellafield. In contrast, the inventory of ^{137}Cs in sub-tidal sediments of the whole Irish Sea in 1988 was 1500 TBq, reducing to 960 TBq by 1995 [147], and it has been estimated that only about 40 TBq of ^{99}Tc are still associated with the sediments of the Irish Sea, down to a depth of 50 cm [148, 149].

There remains some shortfall between the estimated inventories and the reported, decay-corrected discharge. Although part of this can be accounted for by the total loss through the North Channel, it is likely that a significant proportion of the unaccounted fraction is situated in sub-tidal and intertidal sandy deposits below the sampling depth achieved in most published studies [144, 149]. Consequently, more recent publications have updated the radionuclide budget estimation for the eastern Irish Sea [150].

Plutonium in the sediments is reported to be complexed with high-molecular-weight humic and low-molecular-weight fulvic compounds [103, 151]. There appears to be evidence for the reduction of Pu(V) to Pu(III) in anoxic waters, whereupon this reduced plutonium becomes complexed with DOC/COC in the form of humic and fulvic acids. This is a naturally slow reaction in pure water, but its speed increases rapidly in the presence of sediments, in proportion to sediment concentration [103]. Particle surfaces also have the capacity to change the oxidation state of the adsorbed species [152]. Such processes have been verified experimentally in fine sediments under different lighting [145, 153] and in sandy muds collected from the mud-patch near the Cumbrian coast in the vicinity of Sellafield [133].

^{137}Cs uptake kinetics has been studied in the laboratory, demonstrating that uptake by suspended matter occurs immediately after introduction of the isotope and attains equilibrium within a few hours [154]. It would appear that biotite and organic carbon are the most important factors controlling ^{137}Cs fixation onto sediments [155]. In the arctic, ^{137}Cs in surface sediments also correlates with total organic carbon content and is stronger in finer sediment [156].

In contrast with plutonium, americium, and radiocesium, uptake of ^{99}Tc from seawater onto sediment is minimal, and likely to occur only onto anoxic sediments where it may be biogeochemically reduced from $^{99}\mathrm{Tc(VII)O_4}^-$ to insoluble Tc(IV). Conversely, Tc uptake is unlikely in oxidizing sediments low in organic matter [157].

Redissolution from Sediments

The realization that sediments can act not just as a sink but also a potential source of contaminants to the overlying waters has major implications for the assessment of the environmental impact of radioactive discharges into the marine environment [64, 65, 81].

Thus, redissolution rates constitute an important source of information.

The rate of redissolution from sediments is relatively fast for radiocesium, with a predicted desorption rate constant for suspended particles of 1.2 days^{-1} [105]. However this rate is much slower for plutonium and americium [49]. Overall, the "halving time" (time taken for the radionuclide inventory to drop by 50% by considering radioactive decay and rate of removal) is of the order of 10 years for ^{137}Cs, 100 years for ^{239,240}Pu, and 1,000 years ^{241}Am [64–66, 133, 144, 145, 147, 158]. Hence the actinides will persist in the sediments for a considerable time, and in the case of the Irish Sea the water flowing out of the area is one of the sources of transuranics to the shelf seas of Northwest Europe and beyond [81].

That remobilization of plutonium from the sediments of the northeastern Irish Sea is the predominant source for plutonium in this region over the last 2 decades or so has been amply verified [45, 50, 65, 66, 145, 158–160]. An evaluation of the time scales involved in the remobilization of radionuclides from this mud patch (i.e., hold-up/residence times and transport times) is a complex exercise, but it confirms that there is a "hold-up" of plutonium within the eastern Irish Sea sediments of at least a decade [50, 63].

According to an experimental study, the proportions of plutonium and americium available for redissolution from seabed and suspended sediment are rather low at 0.02–1% [133]. This is possibly because plutonium in sediments is predominantly associated with strongly bound sesquioxide and organic complex fractions, while americium is associated mainly with the organic complex fraction, with a significant fraction in carbonate form [133]. The reaction half-time for Pu(V) species to obtain "equilibrium" between solid and liquid phases in sandy sediments is quite short, about 8 h [133]. According to the same study, the process can be accelerated by stirring, which releases 2.5 times more plutonium and americium into seawater, indicating the importance disturbance of the seabed may have on redissolution.

It has also been verified that ^{137}Cs is also desorbing from Irish Sea sediments as a consequence of the general reduction in seawater concentrations [85, 147]. A part of the radiocesium discharged from Sellafield made its way and is probably still to be found in the Arctic Ocean, between Canada and Greenland, with a recorded transit time from Sellafield to Baffin Bay of approximately 8 years [161]. This is not surprising, since monovalent Cs^{+} is less strongly adsorbed than plutonium and americium and therefore it is more easily transported by tides and currents. In poor circulation areas, radiocesium can be remobilized by ion-exchange displacement by cations such as $NH_4{}^{+}$, Fe^{+2}, and Mn^{+2} released under anaerobic conditions [162].

Desorption of ^{99}Tc from contaminated sediment into uncontaminated seawater appears to be highly variable depending on sediment type. Redissolution from oxidizing sediment is rapid, whereas for anoxic sediment it is very slow [157]. In the Irish Sea, where much of the ^{99}Tc discharged appears to be in a truly dissolved form [36], the impact of ^{99}Tc remobilization from Irish Sea sediments upon levels in the water column is thought to be minimal [148]. This is a consequence of the fact that these sediments are mildly reducing, while the overlaying water remains well oxidized during the year [146, 163].

Influence of Bioturbation

Important evidence has been acquired, indicating that bioturbation of radionuclides stored within the sediments is significant, even at some depth [164]. This is important because biological and physical mixing in the upper layers of the sediment column by processes such as bioturbation or gas leakage can disturb the stratigraphic record [165, 166], effectively burying "older" sediment under "younger" deposits [25, 66, 84, 164]. Variations in concentration by as much as a factor of 100 have been observed within a single core [167], making the vertical distribution of sediments notoriously difficult to predict.

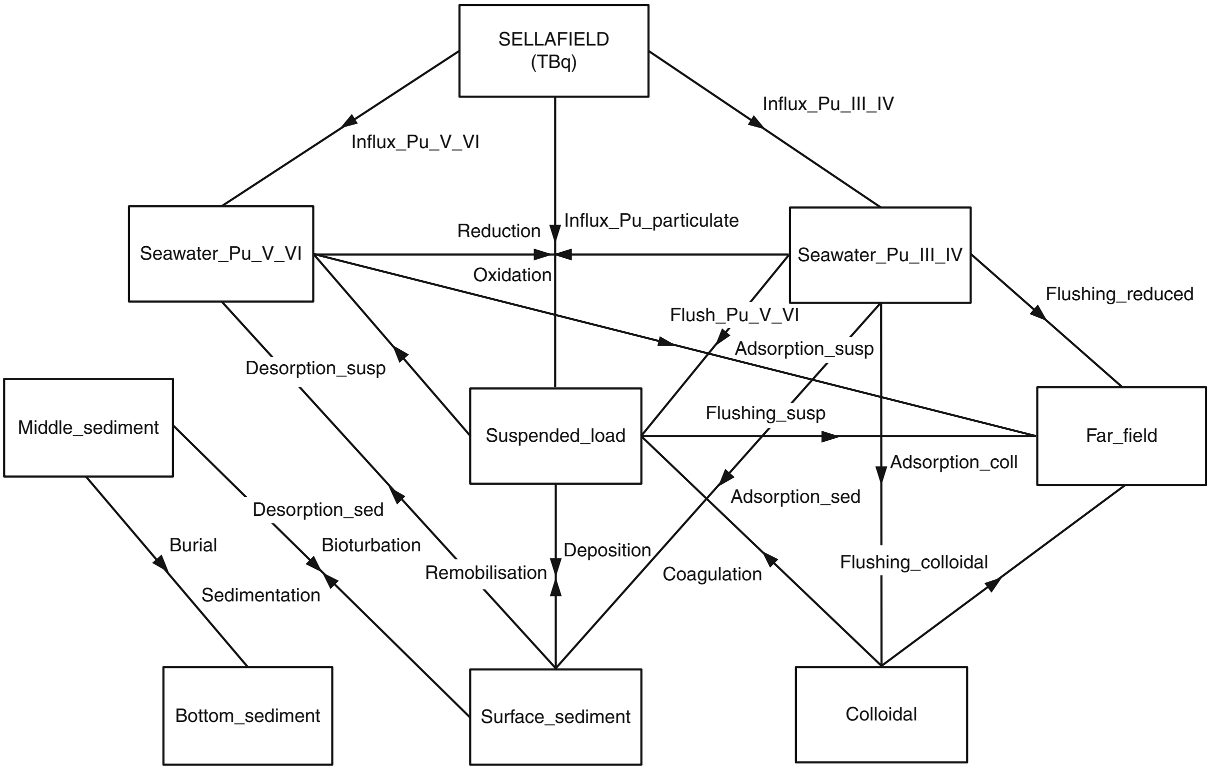
Experimental evidence points to the echurian worm *Maxmuelleria lankesteri* as the main agent responsible for altering the vertical profiles of plutonium and americium in the sediment column of the eastern Irish Sea. The ragworm *Nereis diversicolor* also plays a role in the redistribution of particle-bound radionuclides deposited at the sediment-water interface, with the ability to redistribute ~35% of ^{137}Cs deposited on the sediment surface to ~10 cm depth over a period of just 40 days [168]. The depth distributions of ^{239,240}Pu in deep waters of the Northeast Atlantic Ocean exhibit pronounced subsurface maxima caused by sediment reworking by benthic fauna such as *Sipunculida* worms [169]. These alterations are produced by the organisms feeding on surface sediment and depositing fecal pellets within excavated burrows. The extensive and heterogeneous bioturbation to which sediments are subject, is, undoubtedly, responsible for the variability observed in the vertical profiles of various radio elements [25, 48].

Modeling Studies

It is clear from the foregoing that the sediments of regions like the Irish Sea will continue to be a significant source of plutonium and americium for many centuries [65, 145]. Hence, it is important to model the uptake and redissolution process, predicting the fate of these sediment-locked inventories. The simplest approach is to assume that the sediment is in equilibrium with the surrounding water on an annual basis, applying the concept of mean time of availability, i.e., the average time for which the radionuclide is effectively available to the environment [170]. With this simple method it is possible to reconstruct the deposition history of some radionuclides to a first approximation.

A better approach is to apply compartmental modeling [61, 85, 171, 172]. Irish Sea modeling results confirm the decline of ^{137}Cs and Pu concentrations in water and surface sediments as a result of remobilization and transfer further afield, albeit for the latter radionuclide the process would occur at a much slower rate. The same models predict that the plutonium inventory in deep sediment has not yet peaked, and that this would not be expected to occur for thousands of years.

A process-based model for representing the partitioning of dissolved, colloidal, particulate, and sediment fractions of plutonium and radiocesium has recently been developed [49] (Fig. 3). This model



Radioactivity in the Marine Environment. Figure 3

Schematic of speciation model for plutonium and radiocesium [49]. For clarity, radioactive decay is not represented represents basic biogeochemical processes including a three-layer representation of the seabed, based on previous studies [92, 93]. It confirms the experimental observation that, although in the Cumbrian coastal area near the Sellafield outfall the majority of plutonium in the water column is in a (reduced) particulate form, the majority of the water-borne fraction is in the oxidized state group Pu (V, VI). The model also confirms that most of the plutonium and radiocesium inventories are presently at a depth between 0.1 and 2 m, with the particulate flux from sedimentary deposits to the water column being the main source term for plutonium in the area [85]. Most importantly, the model also predicts the consolidation of the bottom layer inventory over a period of thousands of years.

Behavior of Radionuclides in Estuaries

The environmental behavior of radionuclides depends not only on the nature of the release or source term, but also on the geochemical features of the receiving environment. Considerable variations in the physicochemical behavior of the radionuclides are to be expected in coastal areas, estuaries being a case in point. These behavioral variations are compounded by the fact that the input of inorganic particles, such as clays and metal oxides, can induce important local changes in water chemistry when first entering the sea via estuaries.

There are reports of a wide range of K_d's in tidal estuaries, illustrating the nonequilibrium nature of such systems [51, 173]. Estuaries are highly dynamic environments in which there is a substantial influx of riverine waters, so salinity levels are low. This is accompanied by pH changes and the considerable presence of colloids and suspended particulate matter, competing for the uptake of environmentally available, particle-seeking radionuclides.

In zones where bottom circulation is heavily restricted, such as fjords, there will be a significant depletion in dissolved oxygen levels due to the oxidation of organic matter in bottom waters which are thus converted into an anoxic environment. This results in the reduction of oxidized Pu species reaching the anoxic zone, as observed experimentally in the Framvaren and Hellvik fjords in Norway [174]. Another special case is regions which are shallow and warm by nature, whereupon calcium carbonate tends to form precipitates, which may induce the co-precipitation of radionuclides.

For radionuclides such as plutonium, a redistribution between the dissolved and particulate phase occurs as a consequence of major salinity changes. K_d values can increase substantially when salt concentrations decrease [175]. Consequently, the affinity of particle-seeking radiouclides for suspended sediments is enhanced in these low-salinity environments. Conversely, it has been proposed that small nanoparticles, or colloids, can be important remobilization agents at high salinity [176, 177].

Two cases in point are the Esk Estuary, UK, situated at close proximity to the Sellafield outfall, and the Seine Estuary, which receives an input of plutonium from the Cap de La Hague reprocessing plant. In the Esk Estuary, activity versus salinity plots show higher activities in the dissolved phase below a salinity of 18‰, probably due to the rapid desorption of a labile form of plutonium from suspended particles and sediment in the low salinity in the waters of the estuary [51, 102, 178–180]. The process seems to be associated with a rapid exchange reaction involving competition between plutonium species, protons, and major cations for available solid surfaces [179]. Additional evidence that this is the case comes from measurements of the isotopic quotient, ^{238}Pu/^{239,240}Pu. It has been found that the average quotient in high salinity dissolved fractions is characteristic of contemporaneous discharges, while in the particulate fraction it is characteristic of historical discharges. This has been attributed to the admixture of reworked older estuarine sediment with the original particulate input [51].

Interestingly, the mixing behavior of plutonium in the Esk Estuary at low salinities is the reverse of that observed in the Seine Estuary, where it has been reported that plutonium is lost from the dissolved phase over a wide range of salinities [181]. Suggestions as to the mechanism/s controlling behavior in the Seine Estuary include scavenging of plutonium from solution by reworked uncontaminated estuarine sediments and removal of plutonium from the freshwater input via coagulation with humic acid and iron. This exemplifies the environmental variability of estuaries and the need to understand the processes affecting radionuclide cycling in particular environments.

Other factors besides salinity changes may affect the behavior of radionuclides in estuaries. As discussed previously, enhanced levels of DOC can lead to relatively low plutonium K_d's due to favorable competition of colloids versus suspended particulates for adsorption sites [182]. Differences in the pH of river water compared with seawater may result in particle-seeking radionuclides (chiefly plutonium) arriving from the open sea to redissolve [179]. Another study appears to confirm that, in intertidal sediments, plutonium speciation is influenced by pH changes within estuarine waters [183]. Biological recycling in estuarine intertidal sediments can lead to additional variation through formation of organoliths with a relatively high radionuclide content, following a seasonal pattern [184].

The existence of "hot particles" in estuaries located near reprocessing plant outfalls, known to persist in the environment for months before dissolving [185], can further complicate the picture by increasing the apparent K_d's for particle-seeking radionuclides, such as plutonium and americium [88, 186]. For these reasons, the behavior of radionuclides in the waters of estuaries is difficult to study.

Cycling in the Foodweb

It is clear from previous discussions that life needs to be regarded as an active agent of radionuclide cycling in the marine environment. There is significant biomass in coastal areas, from phytoplankton and zooplankton at the base of the food chain to the higher vertebrates. Radionuclides can be removed from the marine environment by aquatic biota, either by sorption from seawater or by ingestion/egestion of food and sediment. Some species have a high affinity for particular radionuclides, and so they are regarded as biological markers (or bioindicator species) for radiological contamination. They are important aids to help map the transport and distribution of both conservative and nonconservative radionuclides in the marine environment.

Concentration Factors

To express uptake and accumulation of radionuclides by aquatic biota, concentration factors (CFs) are generally used. The concentration factor is the ratio of radionuclide concentration in biological tissue to that in the surrounding water environment:

$$\mathrm{CF} = \frac{\text{Radionuclide concentration in biota } (\mathrm{Bq\ kg^{-1}} \text{ wet weight})}{\text{Radionuclide concentration in water } (\mathrm{Bq\ m^{-3}})}$$

A summary of concentration factors and half-times of elimination ($T_{B1/2}$) for some marine organisms, based on two recent surveys [187, 188], is given in Table 2.

Certain species of brown seaweed (e.g., *Fucus vesiculosus*) have a high concentration capacity for ^{99}Tc. For example, in the Irish Sea, CFs for *Fucus* in excess of 1×10^5 have been recorded [190]. Other species of brown seaweed, such as *Fucus serratus*, also exhibit high concentration factors for technetium, radiocesium, and the transuranics [232, 251–256], leading to their use as bio-indicators. Uptake of ^{99}Tc by the common lobster and the Norwegian lobster is also very significant [199, 200].

Presently, interest on ^{131}I in coastal environments has risen, due to the Fukushima accident. The highest affinity of this radionuclide is for brown seaweed, zooplankton, and phytoplankton, with CFs in the order of 15, 3, and 0.6 $\mathrm{m^3\ kg^{-1}}$, respectively. In molluscs ^{131}I has an affinity for the byssus, operculum, and shell of the molluscs [257, 258]. Iodine $T_{B1/2}$s are in the order of 1–5 days for these species, and a few tens of days for fish and molluscs [187].

Independently from the degree of conservativeness, different elements are incorporated into biota and subsequently depurated at different rates, depending on their chemical and biochemical behavior inside the organism. This means that conservative elements will have low K_d's, but not necessarily low concentration factors or $T_{B1/2}$s in biota. This makes analogue generalizations with regards to biological behavior difficult.

Limitations of the Concentration Factor Approach

The data from Table 2 reflects several decades of work deriving concentration factors for input into models, recognizing the state of equilibrium for the best use of these data in many situations. However, there are limitations to the equilibrium approach. CFs are calculated by comparison of organism activities, which are time integrated, with an average of several

Radioactivity in the Marine Environment. Table 2 Concentration factors (m^3 kg^{-1}) and biological half-lives (d) for selected marine organisms [187, 188]

Species	Data	Tc	Cs	Pu	Am	Source
Cod (*Gadus morhua*)	CF $T_{B1/2}$	2.4E-02 N/A	8.0E-02 N.A.	4.0E-02 N/A	5.0E-02 N/A	Tc: [88, 89, 189, 190]; Cs: [88, 89, 191, 192]; Pu & Am: [88, 89]
Plaice (*Pleuronectes platessa*)	CF $T_{B1/2}$	1.9E-02 3.5E + 01	2.9E-02 6.5E + 01	4.0E-02 N/A	5.0E-02 N/A	Tc: [88, 89, 193]; Cs: [88, 89, 192, 194, 195]; Pu & Am: [88, 89, 196]
Lobster (*Homarus gammarus*)	CF $T_{B1/2}$	3.0E + 00 1.6E + 02	2.2E-02 N/A	3.0E-01 N/A	5.0E-01 N/A	Tc: [88, 89, 140, 193, 197–204]; Cs: [88, 89, 205, 206]; Pu: [88, 89, 198, 200, 202, 203, 207–210]; Am: [88, 89, 198, 202, 206, 208, 211, 212]
Crab (*Cancer pagurus*)	CF $T_{B1/2}$	1.6E-02 6.3E + 01	3.0E-02 N/A	3.0E-01 N/A	5.0E-01 N/A	Tc: [88, 89, 197, 201, 213]; Cs: [88, 89]; Pu: [208]; Am: [88, 89, 208]
Winkle (*Littorina littorea*)	CF $T_{B1/2}$	2.2E + 00 1.4E + 02	3.8E-02 8.6E-01	1.9E + 00 1.2E + 02	7.3E-01 1.8E + 02	Tc: [88, 89, 200, 214, 215]; Cs: [45, 88, 89, 216–220]; Pu & Am: [45, 88, 89, 216, 217, 219–223]
Cockle (*Cardium edule*)	CF $T_{B1/2}$	1.7E + 00 N/A	3.0E-02 N/A	1.4E-01 6.5E + 00	2.4E-01 2.0E + 01	Tc: [88, 89, 140]; Cs: [88, 89]; Pu & Am: [88, 89, 223, 224]
Mussel (*Mytilus edulis*)	CF $T_{B1/2}$	9.8E-01 N/A	2.1E-02 1.8E + 01	1.9E + 00 7.1E + 02	1.2E + 00 3.0E + 02	Tc: [88, 89, 200]; Cs: [45, 88, 89, 217, 219, 220, 223, 225]; Pu: [45, 88, 89, 217, 219, 220, 222, 223, 226]; Am: [45, 88, 89, 219, 220, 222, 223, 226, 227]
Brown seaweed (*Fucus vesiculosus*)	CF $T_{B1/2}$	2.2E + 01 1.3E + 02	5.4E-02 5.4E + 01	3.0E + 00 N/A	4.3E + 00 1.4E + 02	Tc: [88, 89, 140, 200, 215, 228, 229]; Cs: [45, 88, 89, 140, 218–220, 230–234]; Pu: [45, 88, 89, 218–220, 235]; Am: [45, 88, 89, 219, 220, 232, 236, 237]
Zooplankton	CF $T_{B1/2}$	5.5E-02 N/A	3.0E-02 1.3E + 01	5.5E-01 N/A	1.1E + 00 3.4E + 01	Tc: [88, 89, 238]; Cs: [88, 89, 239]; Pu: [88, 89, 113, 238, 240, 241]; Am: [88, 89, 240–243]
Phytoplankton	CF $T_{B1/2}$	6.3E-03 2.1E + 00	2.0E-02 2.6E + 01	2.8E + 02 6.1E + 00	2.2E + 02 1.3E + 01	Tc: [88, 89, 238, 244, 245]; Cs: [88, 89, 246]; Pu: [88, 89, 241, 247, 248]; Am: [241, 247, 249, 250]

instantaneous seawater results (or even a single result). This problem can be compounded by the large range of seawater activity typical of many coastal inputs [215].

Although aquatic organisms equilibrate more rapidly than terrestrial organisms with radionuclide concentrations in the surrounding media [259], there are cases in which equilibrium cannot be assumed, the timeframe of interest being the limiting factor. For example, after a short-term pulsed release of Tc activity, the activity concentration in lobsters along the dispersion path begins to increase gradually. Lobsters have a long biological half-life of elimination for technetium (66 days for uptake from food and 200–300 days for uptake from seawater) [203]. However, technetium is soluble in seawater and will clear quickly from the area where lobsters live, aided by tides and currents, so its concentration in seawater is likely to decrease sharply within a few days. If a lobster has been sampled within days of the discharge, it may appear to have an anomalously high "concentration factor" because it still retains the technetium that it absorbed, while the initial Tc pulse has cleared from the surrounding water.

In other words, if equilibrium has not been reached, the measured CF will only reflect the concentration ratio at a given point in time rather than the steady-state CF that can be measured in the laboratory under controlled conditions. This is why uptake of radionuclides such as ^{99}Tc in the field has occasionally been observed to exceed that which has been reported in laboratory studies [260].

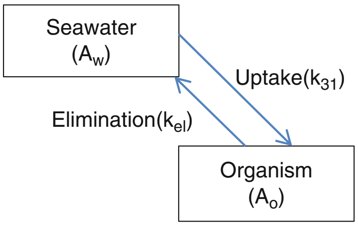
Despite the above shortcomings, if used correctly, concentration factors can be used to effectively illustrate trends and differences between species. If the timeframe of interest is long (e.g., years or decades of planned authorized discharges, involving continuous releases or gradual changes in discharge concentrations) then the CF approach may well suffice. If the timeframe of the assessment is short (e.g., hours, days, or weeks, such as in emergency or unplanned release scenarios involving abrupt changes in discharge concentrations) dynamic models of the radionuclide transfer to biota are a better assessment tool. This is especially true for organisms that respond slowly to a change in ambient radioactivity concentration.

Bio-Accumulation as a Dynamic Process

The assumption of equilibrium in marine ecosystems can lead occasionally to erroneous estimations of radionuclide transfer to biota. As an illustration, Jackson et al. [215] made the observation that concentrations of ^{99}Tc in seawater off the Cumbrian coast after episodic discharges of ^{99}Tc from Sellafield decreased relatively rapidly, while those in seaweed (*Fucus vesiculosus*) and winkles (*Littorina littorea*) did not. This demonstrates the limitations of using an equilibrium factor to predict radionuclide uptake under discharges of an intermittent nature.

Several methods have been developed to derive activity concentrations in biota under nonequilibrium conditions. One aspect of the dynamic process approach is that these models can very quickly become data "hungry," requiring many terms and parameters that are difficult to measure. However, relatively simple biokinetic models based on first-order exchange kinetics between the medium and the organism are a good improvement. These models use relatively few data: the biological half-life of elimination and the CF (which, as an equilibrium parameter, still has a role to play in the calculations) to deduce the exchange rates between the medium and the organism [188, 261–266].

The simplest approach is to assume a two-compartment first-order kinetic model with constant rates of uptake and excretion from water (k_u and k_{el}), as shown in Fig. 4. With such a simple model, it is possible to deduce the CF for the steady-state as a function of the model rate constants:



Radioactivity in the Marine Environment. Figure 4 Simple 2-compartment biokinetic model

$$\mathrm{CF} = \frac{k_u}{k_{el}} \frac{V}{m}$$

where m and V are the mass of the organism (kg) and the volume of water (L) in which the organism is immersed. The kinetic rates k_u and k_{el} can be simply written in terms of the biological half-life of elimination after uptake from water ($T_{B1/2}$), which is the length of time required for all combined physiological processes to cause the loss of half of the bio-accumulated radionuclide from an organism. Hence:

$$k_{el} = \frac{\ln 2}{T_{B1/2}} \quad \text{and} \quad k_u = \frac{\ln 2}{T_{B1/2}} \frac{m}{V} \mathrm{CF}$$

This simple way to calculate the uptake and release rates allows an estimation of the response time for biota exposed to a discharge of radioactivity. For example, it takes some 1–1.5 years for activity in winkles (*Littorina littorea*) and lobsters (*Homarus gammarus*) to peak, following an increase in generalized ^{99}Tc levels in seawater [188, 260].

The above explanation is somewhat simplified because radionuclide turnover by marine organisms can be multiphasic, i.e., with an initial fast release followed by a slower, longer-term release, with a point in time marking the transition from one phase to another. There are many instances of this, such as the biphasic release of Tc in seaweed with $T_{B1/2}$s of ~1 and ~100 days, respectively [140, 200, 229]; Tc and Pu in European lobsters [193, 201, 209, 260]; ^{241}Am and ^{237}Pu from contaminated mussels (*Mytilus galloprovincialis*) [226], or the typical multiphasic release curve representing the depuration of ^{131}I from *Littorina* *littorea* [267–269] (Fig. 5), to quote but a few examples. Details of other occurrences are given elsewhere [187, 188].

Recent modeling studies have attempted to represent organ assimilation, direct excretion, and delayed excretion, capturing multiphasic release in the process. A typical improvement with respect to Fig. 2 involves adding more compartments with linear transfer rates between different organs [260, 270, 271]. An adequate compromise is a three-compartment model which makes the working simplification that marine organisms absorb radionuclides mainly from the surrounding seawater, allowing to relate their radioactivity concentration to their environment (Fig. 6).

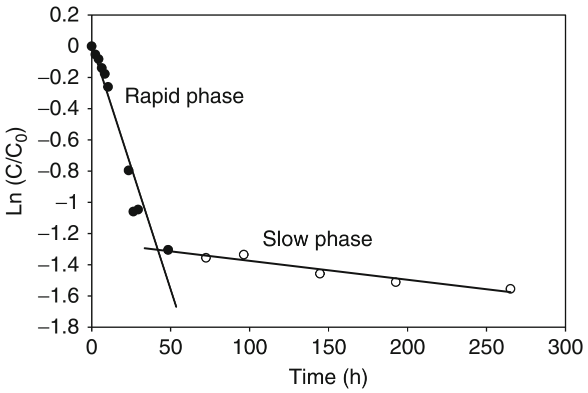
The above model includes direct exchange with seawater of specific activity A_w through both a fast compartment of activity A_f and a slow compartment of activity A_s:

$$\frac{dA_w}{dt} = k_{el}^{f}A_f + k_{el}^{s}A_s - (k_u^f + k_u^s)A_w$$

$$\frac{dA_f}{dt} = k_u^f A_w - k_{el}^{f}A_f; \frac{dA_s}{dt} = k_u^s A_w - k_{el}^{s}A_s$$

Such a model incorporates biphasic release, being sufficiently simple to be solved analytically without recourse to numerical computation. The model requires relatively few parameters, rendering the approach practical for use in radiological assessments [188].

The biological half-life of a radionuclide in an organism is not always constant. It can be affected by factors such as: the fed or starved state, resulting in slower turnover for the latter (observed for ^{99}Tc in gastropods) [214]; temperature, with higher depuration rates at higher values (observed for ^{99}Tc in shrimps [272] and Pu in lobsters [209]); the physical and chemical speciation of the radionuclide; physiological characteristics (size, age, stage of development, sex, habitat); and certain physiological processes associated with molting in crustaceans, which tend to temporarily increase the uptake rate. The latter has been observed for Tc [197, 204] and Pu [209] in crabs and lobsters.



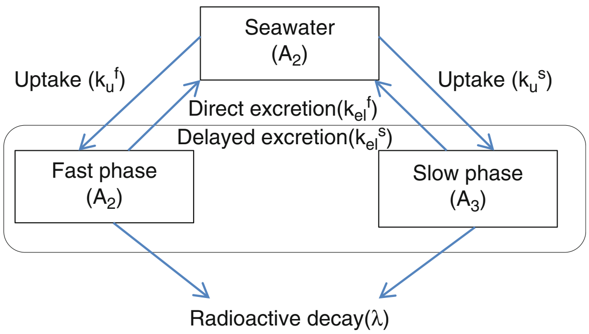
Radioactivity in the Marine Environment. Figure 5 Multicomponent release curve for ^{131}I depuration from *Littorina littorea* [268]

Allometry

The introduction of the biological half-life as an additional transfer parameter is not without its difficulties, as data are not always available. Allometric approaches provide some guidance when estimating transfer rates for use in models [261, 273–275], as well as providing a tool for data interpretation [276]. It has been observed that most metabolic parameters, including basal metabolic rates, ingestion rates, biological half times, etc., are proportional to a simple power function of organism mass. This was first formulated by Kleiber [277, 278], who predicted that the metabolic rate B_r of an organism of mass M is proportional to a quartile power of the mass M:

$$B_r = aM^b; \quad b = 0.75$$

Various radionuclides in marine biota exhibit allometric relationships with respect to the CF and the $T_{B1/2}$, relating to metabolism [279]. A recent study found



Radioactivity in the Marine Environment. Figure 6 Basis of a dynamic model for the transfer of radionuclides to marine biota [188] power-like allometric associations between the biological half-life (^{137}Cs, ^{99}Tc) or the concentration factor (Pu, ^{241}Am) on the one side, and size in marine biota [187, 280], as shown in Table 3.

This same study found an association of the independent term of the allometric power function (b) with the K_d, at least for particle-seeking lanthanides and actinides.

It is not surprising that, on average, the CF correlates with mass to the power of about −0.26, since the CF relates to ingestion rate per unit mass, and the ingestion rate is proportional to the metabolic rate, hence CF ~ $M^{0.75} \times M^{-1}$or $M^{-0.25}$. The relationship between $T_{B1/2}$ and metabolic rate is more difficult to explain, but it can be derived using a dynamic model developed for the prediction of radionuclide activity concentrations in Arctic marine species (ECOMOD) [281]. This approach considers that, for a growing organism of total mass $M(t)$, the radionuclide activity concentration y (Bq kg^{-1}, fresh weight) in a given tissue changes according to:

$$\frac{dy}{dt} = -\left(\lambda + \varepsilon_a \frac{B_r}{M} + \frac{1}{M}\frac{dM}{dt}\right) y + \frac{Q_1^A}{Q_0^A}\left(\frac{1}{M}\frac{dM}{dt} + \varepsilon_a \frac{B_r}{M}\right) X(t)$$

where λ is the radioactive decay constant (s^{-1}); ε_a is a proportionality coefficient between the rate of biological loss of a radionuclide from a deposit tissue and the general metabolic rate of the organism; B_r is the metabolic rate (kg s^{-1}); $X(t)$ is the radionuclide activity concentration in food (Bq kg^{-1}, fresh weight) or in water (Bq L^{-1}) when bio-assimilation occurs directly from water at time t, and Q_1^A,Q_0^A are the stable element concentrations in a given tissue and in food or water (mg kg^{-1}), respectively.

For the depuration case, in which $X(t) = 0$, if assuming that the mass of the organism is constant and neglecting radioactive decay, the above equation returns $y(t) = y(0)e^{-\varepsilon_a \frac{B_r}{M} t}$. According to this model, the concentration halving time, or biological half-life of elimination, is $T_{B1/2} = \frac{\ln(2)M}{\varepsilon_a B_r}$. Hence, $T_{B1/2} \sim M \times M^{-0.75} \sim M^{0.25}$, consistent with the prediction that turnover times generally change with a quartile power of the mass [280, 282].

Food Chain Transfer and Biomagnification

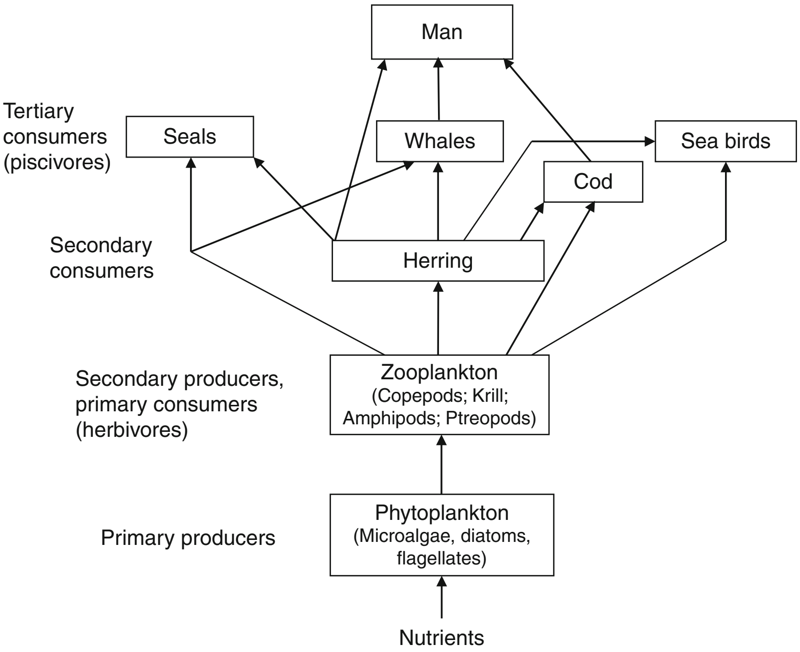
The transfer of radionuclides through the marine food chain opens the potential for biomagnification, i.e, the build-up of concentration of certain contaminants, such as toxic metals and persistent halogenated organic compounds, in the bodies of organisms at higher trophic levels. As smaller organisms are ingested by larger ones within the food chain (Fig. 7), biomagnification could theoretically result in higher concentrations of the substance than would be expected if water were the only exposure mechanism.

The potential problem caused by biomagnification has been relatively under-investigated. A modeling study was carried out in the Arctic under the INCO-Copernicus EPIC program [281]. According to this study, ^{137}Cs biomagnification could be occurring for the lower trophic levels (e.g., benthic food chains), similar to what has been observed for other elements such as Cd, methyl-mercury (MeHg), and Po in a chain

Radioactivity in the Marine Environment. Table 3 Allometric relationships for CF and $T_{B1/2}$ in marine biota [280]

	$\log_{10}$[CF(m^3 kg^{-1})]								$\log_{10}$[$T_{B1/2}$(d)]			
Parameter	Ru	Ce	Pm/Eu	Ra	Th	Pu	Am	Cm	Tc	Cs	Pu	Am
Log_{10}a	−1.54	−0.33	0.19	−1.07	−0.25	−0.68	−0.60	−0.44	1.99	1.73	2.80	2.40
b	−0.46	−0.25	−0.18	−0.11	−0.27	−0.30	−0.28	−0.27	0.15	0.17	0.20	0.13
r^2	0.75	0.72	0.72	0.79	0.89	0.80	0.78	0.85	0.72	0.92	0.91	0.76
SE(Log_{10}a)	0.37	0.22	0.17	0.08	0.14	0.22	0.21	0.17	0.20	0.08	0.39	0.22
SE(b)	0.10	0.06	0.04	0.02	0.04	0.06	0.06	0.04	0.05	0.04	0.06	0.04
p-value	0.002	0.004	0.004	0.001	0.0002	0.001	0.002	0.0004	0.03	0.04	0.2	0.05
N	9	9	9	9	9	9	9	9	6	4	3	5

Note: SE, Standard error of the mean


Man
Tertiary consumers (piscivores)
Seals
Whales
Sea birds
Cod
Secondary consumers
Herring
Secondary producers, primary consumers (herbivores)
Zooplankton (Copepods; Krill; Amphipods; Ptreopods)
Primary producers
Phytoplankton (Microalgae, diatoms, flagellates)
Nutrients

Radioactivity in the Marine Environment. Figure 7
Example of a simple marine food chain containing the basic trophic levels. Adapted from [283]

from plankton to killifish [284]. However, this study suggests that the same is not happening at the highest trophic levels. For ^{239}Pu, transfer to successively higher trophic levels is shown to be low – there is a fall of several orders of magnitude between primary producers, represented by phytoplankton, and higher trophic levels, e.g., cod. Further investigations on polar foodwebs in the waters of the Barents Sea also concluded that transport of ^{137}Cs upward through the food chain does not take place [285]. On the basis of this limited information, it is unlikely that radionuclides from the Fukushima accident will be magnified up in marine food chains, except possibly for the lower trophic levels.

Biodiminution of metals in the planktonic food chain (phytoplankton to copepods to fish) has been attributed to the outflow of metals by copepods and the very low capacity of marine fish to assimilate metals [286]. However, there may be some exceptions to this rule, such as whales, whose diet is based on zooplankton (which have a relatively high concentration capacity for plutonium) [281]. Nevertheless, the potential for the biomagnification of many radionuclides appears to be low. This is likely to be due to biological discrimination against uptake of elements having limited biochemical usefulness, an argument which may not apply to chemical analogues which are incorporated as nutrients into living organisms.

Radiological Implications of the Marine Environment

Human Radiological Exposure

The worldwide fallout from nuclear weapons tests led to the transfer of radionuclides via the marine food chain and an internal dose rate to man through ingestion of the species responsible for human exposure: fish, shellfish, and seaweed, as well as external exposure to beaches and fishing nets. It has been calculated that the corresponding dose rates are very low compared with the dose rates received by general populations from the terrestrial environment. For general populations depending heavily on fish consumption, dose rates could have been at most in the order of 4 μ-sieverts per year (μSv y^{-1}) to the whole body [132].

Radiologically, the seas most affected by the Chernobyl accident were the Baltic and Black Seas, where enhanced ^{137}Cs was observed in aquatic food. However, the dose rates to the public from ingestion of ^{137}Cs from Chernobyl in aquatic food have been estimated to be low, at least an order of magnitude lower than those due to natural ^{210}Po [287]. The highest dose from the Chernobyl accident was delivered to the critical group in the Baltic Sea region, who received a maximum dose of 0.08 mSv during 1986 [288].

Exposures to routine discharges from the nuclear industry to coastal sites have generally given rise to dose rates of a few tens of μSv per year for the majority of power stations. Some examples in the 1985–2000 period are: 10 μSv y^{-1} in Bradwell, UK, arising from the consumption of ^{137}Cs in fish (1996); 10 μSv y^{-1} in Heysham, UK, arising from external gamma exposure to ^{60}Co (1991); 70 μSv y^{-1} for Paluel, France, arising from consumption of crustaceans and molluscs contaminated with ^{110m}Ag in 1987; and 15 μSv y^{-1} for Dounreay power station, UK, from ^{137}Cs in fish in 1987 [289]. An up-to-date account of exposures due to discharges from the UK nuclear industry in 2009 gives total dose rates to the public for liquid releases from the site of <5, 49 and 11 μSv y^{-1} for Bradwell, Heysham and Douneray, respectively [290]. The dominant contribution is due to gamma dose rates to adult occupants over sediment (only Sellafield has an exposure contribution from the consumption of seafood).

The two most important contributions over the time period 1985–2000 in terms of human dose were the reprocessing plants at Sellafield (UK) and Cap de la Hague (France). According to a modeling study, in Cap de la Hague, dose rates dropped steadily from 170 μSv y^{-1} in 1987 to 19 μSv y^{-1} in 1996, being dominated throughout this period by the consumption of molluscs contaminated with ^{106}Ru and ^{241}Pu and external exposure to gamma-rays from ^{60}Co in sediments.

The radiological impact of Sellafield releases to the Irish Sea environment has been studied in detail using monitoring data and habits surveys [4, 23]. According to a thorough review [291], during the 1950s and 1960s, the highest dose rates from Sellafield were received by individuals consuming Cumbrian *Porphyra* and laverbread, with peak dose rates around 1 mSv y^{-1}. A maximum dose rate of 3.5 mSv y^{-1} (arising principally from ^{137}Cs discharges) would have been reached by 1981.

The maximum dose rate of 3.5 mSv y^{-1} for Sellafield has been revised in line with a reassessment of the porphyra/laverbread ingestion pathway. It is now thought that the critical pathway was fish and shellfish consumption, particularly actinides in molluscs, which tend to concentrate transuranics to a considerably greater extent than do crustaceans, in turn accumulating more than fish. Peak exposures to the critical group of about 2 mSv y^{-1} would have been received in the mid-1970s [292].

By the mid-1980s dose rates from Sellafield decreased to about 0.5 mSv y^{-1}, arising mainly from ingestion of Pu and Am. Exposures then decreased to 187 μSv y^{-1} in 1987 and further down to 114 μSv y^{-1} in 1996, in line with reductions in the transuranic discharges [289].

After the enhanced actinide removal plant (EARP) began operations in 1994, ^{99}Tc discharges from Sellafield increased by almost two orders of magnitude, reaching a maximum in 1995. This resulted in relatively elevated critical group dose rates arising from the consumption of Tc in shellfish (mainly lobsters), 50 μSv y^{-1} in 1997 [293]. As the backlog of Tc effluent was processed by the plant and the soluble Tc began to clear from the Sellafield coastal area, these dose rates decreased sharply [294].

Since then, dose rates from Sellafield to all critical groups have continued to decline [81], currently being very low compared to the peak discharge periods of 1995 for ^{99}Tc, 1975 for ^{137}Cs, 1973 for ^{239,240}Pu, and 1974 for ^{241}Am. In 2009, e.g., a combined dose rate of 0.28 mSv y^{-1} for Sellafield and LLWR sites (^{239,240}Pu and ^{241}Am) on the one hand, and the phosphate processing works in Whitehaven (^{210}Po) on the other, was reached [290]. The Pu and Am contribution at the time, majorly attributable to Sellafield, was about 0.15 mSv y^{-1}. The main pathway continues to be adult crustacean and mollusc consumption.

Along the Irish coast, dose rates have been much lower. Starting from 14 μSv y^{-1} in the early 1980s when the Sellafield discharges were close to their peak, exposures have decreased sharply to <0.1 μSv y^{-1} at present, arising mainly from Pu and Am from consumption of fish and shellfish [295]. ^{99}Tc dose rates to Irish typical and heavy seafood consumer groups in the period 1996–1998 were low at 0.061 and 0.24 μSv y^{-1}, respectively [296].

Long-range transports of radionuclides originating from fallout and from reprocessing plants (including remobilization of Pu and Cs from Irish Sea sediments) are still the main sources of man-made radionuclides to the Norwegian marine environment. An additional source was outflow of Chernobyl ^{137}Cs-labeled waters from the Baltic Sea [297]. Using an average consumption rate for a Norwegian individual and average concentration compiled for a suite of radionuclides (including anthopogenic and naturally occurring radionuclides) from Norwegian marine areas, an approximate annual dose rate of 0.1 mSv has been derived from the consumption of fish and other seafood. Some 99% of the dose contribution is from naturally occurring radionuclides [298].

The Marina Med project [22] assessed doses in the Mediterranean region for a generalized critical group with high fish and shellfish consumption rates, using measured seawater concentrations and the known biota concentration factors. It was calculated that the critical group received doses of 0.5 μSv y^{-1} from ^{137}Cs and 540 μSv y^{-1} from ^{210}Po in 1990 [288]. The Mediterranean also received an additional input of ^{137}Cs as a consequence of the Chernobyl accident. Calculations showed that, by that time, the critical Mediterranean group was receiving an integrated effective dose from Chernobyl ^{137}Cs in marine food of about 7.5 μSv y^{-1}, about five times lower than the Black Sea critical group and two times lower than Baltic critical group [23].

In the northwestern Pacific Ocean, the average individual dose rates from seafood ingestion in the early 2000s were estimated to be 0.03 μSv y^{-1} from ^{137}Cs and 9 μSv y^{-1} from ^{210}Po. The calculated annual dose of ^{137}Cs for a hypothetical critical group of heavy fish and shellfish consumers was 3 μSv y^{-1}, while the contribution from ^{210}Po would be 160 μSv y^{-1}[299].

In the Arctic generally, by the late 1990s dose rates to humans from the consumption of seafood were dominated by ^{137}Cs and estimated to be 2.6 μSv y^{-1} due to seafood from the Barents Sea and 4.2 μSv y^{-1} via seafood from the Kara Sea [300]. Meanwhile, in the Antarctic Ocean, the individual dose to a critical group consuming molluscs as a result of NEA dumping activities was modeled to be 0.1 μSv y^{-1}, arising mainly from ^{239}Pu and ^{241}Am [301].

Overall, the various sources of anthropogenic radionuclides in the marine environment, such as global fallout, nuclear weapons testing, releases from nuclear facilities, radioactive waste dumping, the Chernobyl accident and nuclear submarine and aircraft accidents, give individual dose rates from ingestion of marine food which are presently a few μSv y^{-1}. For comparison, the annual dose rate for a heavy consumer of fish and shellfish due to the naturally occurring ^{210}Po would be 150–160 μSv y^{-1} [302, 303].

Given the abundance of dosimetric studies on coastal sources of radioactivity, this summary is not comprehensive. However, it is clear that for most situations dose rates from coastal inputs of radioactivity are well below the annual average dose rate of ~3 mSv received by members of the general population from all sources of radiation, except for external irradiation in a few contaminated spots [304]. The dominant contribution to dose actually derives from natural ^{210}Po in fish and shellfish, compared with which the contribution from anthropogenic ^{137}Cs (mainly originating from nuclear weapons tests and the nuclear industry) is generally lower [305, 306].

In the early 1990s, the IAEA MARDOS study confirmed that the dominant contribution to marine dose rates to man comes from natural ^{210}Po in fish and shellfish and that the contribution of anthropogenic ^{137}Cs (mostly from fallout) is negligible, as are the marine dose rates from other man-made radionuclides occurring in the oceans [307]. This confirms that, taken globally, radiation dose rates to humans from exposure to the marine environment and its foodstuffs have little radiological significance.

Radiological Protection of Non-Human Biota

A particular problem surrounding the assessment of radiation dose to non-human biota is that, unlike for humans, an internationally accepted system for the radiological protection of animals and plants is still under development, although the need for such a system to demonstrate that the environment is protected from ionizing radiation is now broadly recognized [308–310]. The old tenet that if humans are protected biota are also protected is no longer accepted [311, 312]. For example, renewed dumping of low-level radioactive waste to the sea would result in low dose rates to humans due to the high dilution in transit from remote areas, but there could be effects in biota that live at the dumping sites. In another example, the maximum radiation dose rate to the marine mammals in coastal waters around the UK (mainly from radiocesium) is estimated to have been higher than dose rates previously assessed for critical groups of humans living near Sellafield, while the maximum dose rate from plutonium is comparable to the dose rates for humans [313].

As the need to ensure explicit radiological protection of the environment is felt, several national and international bodies have developed or are developing assessment methodologies for nonhuman biota [188, 259, 266, 275, 308, 314–320]. The current state of knowledge underpinning these methodologies is being continually improved and cross-compared [321], but it is generally accepted that prediction of the transfer of radionuclides to organisms (as explained previously) is a major source of uncertainty [322, 323].

Compounded with the above are the challenges posed by the derivation of benchmarks with which to evaluate the scale of radiation doses to nonhuman biota. There is a general consensus on dose rate levels that are unlikely to cause effects to flora and fauna: Brown et al. [324] concluded that only minor effects on biota are to be found for dose rates <100 μGy h^{-1}, while the EC ERICA project, based on a detailed statistical analysis, proposed a benchmark level of 10 μGy h^{-1} [316]. UNSCEAR concluded that dose rates up to 400 μGy h^{-1} to a small proportion of individuals in aquatic populations would not have a detrimental effect at the population level [325]. In their revised recommendations, the International Commission on Radiological Protection (ICRP) proposed a "derived consideration reference level" (DCRL) of 4–40 μGy h^{-1} for the most sensitive reference animals and plants [318]. A generic screening value of 10 μGy h^{-1} seems reasonably robust, on the basis of current knowledge [326], and is within range of natural background doses for aquatic organisms of up to a few tens of μGy h^{-1}.

Nevertheless, knowledge on *what effects may be expected* at radiation dose rates in excess of 10 μGy h^{-1} is currently very sparse, and there is little known in respect of long-term effects to entire populations. This limits the assessment of the implications of exceeding a "screening dose rate" and, given the amount of data that would be required, such limitation is unlikely to be resolved in the near future.

During recent years, a number of studies have been performed to estimate radiation doses to marine biota. Such estimates are generated using purposely developed assessment tools or models, a fact that should be borne in mind given the evolving scientific position in the dose-to-biota assessment field. However, actual experiments to measure doses to biota have also been carried out. For example, in a pioneering study [327] plaice were tagged with dosimeters and released into the eastern Irish Sea environment in an attempt to generate "measured" exposure rates of biota in their natural state.

In the Sellafield zone, dose rates for ^{239}Pu in benthic molluscs were calculated to have been 10^{-2}μGy h^{-1} in the 1950s. They increased to about 7 μGy h^{-1} in the late 1970s, when transuranium discharges were at their peak. By 2005, plutonium doses had diminished to about 0.1 μGy h^{-1}. For ^{99}Tc in lobsters it has been estimated that, in the 1950–1970s, dose rates were less than 10^{-2} μGy h^{-1}. Then, in the late 1970s, they rose to a peak of about 0.1 μGy h^{-1}. Between 1980 and the early 1990s Tc doses decreased to pre-1970s levels. Finally, in the early 1990s, they peaked again to above 0.1 μGy h^{-1}, due to the treatment of EARP plant effluent, before falling again to low levels at present [328].

Another study gives doses of the order of 1.3×10^{-2}, 0.17 and 17 μGy h^{-1} for fish, crustaceans, and molluscs over the period 1986–2001 [329]. With the exception of molluscs, these values are low compared with doses to biota local to the Whitehaven phosphate plant, UK (where enhanced levels of naturally occurring radioactive material were processed until 1992), or 4×10^{-2}, 0.4 and 0.8 μGy h^{-1} for fish, crustaceans and molluscs by the early 1990s [329].

It can be concluded that in the Irish Sea near Sellafield, even during the period of maximum discharges, dose rates to marine biota were below those which would cause any measurable effect at the population level [84, 330], with exposures expected to decline significantly as a result of the dispersion and mixing of water and sediments in the presence of falling discharges [81].

Dose rates to marine biota in the Cap de la Hague coastal area have been calculated to be somewhat lower than those at Sellafield. During the period 1986–2001, for which a major international assessment exercise was performed [329], the most exposed organisms were molluscs and crabs, with beta and gamma dose rates varying between 6×10^{-3} to 6×10^{-1} μGy h^{-1}. For Pu-isotopes the dose rate (weighted by radiation weighting factor) was estimated to be 0.6–1.7 μGy h^{-1} for mussels and 8×10^{-3} to 2×10^{-2} μGy h^{-1} for fish [329]. Subsequently, there has been a general tendency for doses to decrease, with contributions from ^{3}H, ^{14}C, ^{60}Co, ^{90}Sr, radiocesium, Pu, and Am to crustaceans, molluscs, fish, and algae reaching levels not in excess of 5×10^{-2} μGy h^{-1} by 2003 [331].

In the fjords of the Novaya Zemlya archipelago (Abrosimov and Tsivolki Fjords) where some of the world's largest radioactive waste dumping sites are located, radiation dose rates to the local marine biota in 1992–1994 (predominantly from α-emitters) appear to have been small (in the order of 0.2, 0.6, and 3.7 μSv y^{-1} for zooplankton, molluscs, and fish, respectively), with maximum external dose rates up to 3 mSv y^{-1} for molluscs living within small contaminated spots on the seabed [304]. These dose rates would have not been exceeded by biota from the Kara Sea generally.

By the end of the 1990s, dose rates for biota in the Barents Sea, where the main contributors to ^{137}Cs contamination are the European reprocessing plants and Chernobyl fallout draining through the Baltic Sea, were also low at 0.1–0.3 μGy h^{-1}. The factor driving the dose was external exposure from ^{137}Cs in bottom water and sediments [285]. More recent data suggests that, in remote areas of the Barents Sea, doses of 0.8–3 $\times 10^{-4}$ μGy h^{-1} exist for fish, crustaceans and molluscs, with crustaceans dominating [329].

Overall, for the whole of the nuclear industry operating across the OSPAR region, a range of 10^{-4}–10^{-1} μGy h^{-1} was estimated for exposures in macroalgae, crustaceans, and vertebrates by 2005 [332]. Other sources of radioactivity appear to pose lower impact on marine biota. For example, NORM radionuclides in oil and gas rigs in Norway ($^{226/228}$Ra, ^{210}Pb, ^{210}Po) have given rise to doses to fish, molluscs, phytoplankton, and zooplankton of as little as 5×10^{-3} –7×10^{-2} μGy h^{-1} by 1996 [333].

A recent development in studies of impact to nonhuman biota is the incorporation of dynamic transfer of radionuclides as part of the assessment, to address situations in which ambient concentrations change rapidly and $T_{B1/2}$s are protracted, such as in the vicinity of coastal inputs where equilibrium is almost never reached. One such study [328] predicts lower dose rates than calculated with the equilibrium CF approach when ambient activities are rising, and the opposite when they are decreasing. For example, using the Environment Agency (UK) R&D 128 dose to biota assessment tool [259], which is equilibrium based, peak ^{99}Tc dose rates in the Drigg area near Sellafield during the period 1997–1999 were calculated to be 7 μGy h^{-1} for brown seaweed (*Fucus vesiculosus*) and 0.7 μGy h^{-1} for gastropod (*Littorina littorea*). These doses correspond to a ^{99}Tc pulse released around January 25, 1998. An alternative dynamic model generates lower predictions of 3.5 and 0.1 μGy h^{-1}, respectively. This effect should be borne in mind when assessing the radiation dose to marine biota under pulsed discharges of radioactivity.

On the basis of the above, it can be concluded that doses to biota above a level of 10 μGy h^{-1}are hardly achieved under planned exposure situations [334], though the same cannot be concluded for accidents. It is necessary to continue investigating the radiological impact to benthic organisms inhabiting the sediments, because sediments constitute the eventual sink for many radionuclides entering the marine environment and these habitats may be sensitive, due to the presence of sediment-dwelling organisms such as polychaete worms.

Conclusions

It is clear from this overview that a considerable body of knowledge has been accumulated over the years on the behavior of radionuclides discharged into the marine environment, including the transfer processes responsible for their cycling and their impact on wildlife. The present review cannot be exhaustive, given the many and distinguished investigations carried out in this field of research, but it suffices to illustrate the amount of effort that has gone into understanding the physico-chemical speciation, bioavailability, and radiological impact of radionuclides in marine systems, through numerous field and laboratory studies.

From the evidence presented, it can be concluded that, by the complexity of their physical and chemical properties, radionuclides have invaded every component of marine ecosystems: soluble fraction, particulates, colloids, suspended and seabed sediments, and the life forms inhabiting therein. The circulation and fate of these radionuclides is inextricably linked to the marine processes occurring in a uniquely complex environment, and thus radionuclides are very useful as tracers in the study of marine processes.

The key chemical factor determining radionuclide behavior in the marine environment appears to be its oxidation state distribution, i.e., speciation. The key factors affecting horizontal transport are advection and tidal currents, whereas vertical transport is driven by uptake and scavenging by organic and inorganic particulate matter, having a strong capacity to remove radionuclides from the water column. The importance of this process varies considerably depending on the environment, whether it be shallow coastal areas, intermediate depth, or the deep ocean [121, 335]. In shelf areas, radionuclides are removed from the water column more quickly by the deposition of fine grain sediments. In open oceans the key mechanism may be in falling fecal matter from plankton.

The ultimate fate of radionuclides in the environment is to disperse globally across the oceans, with a certain fraction of the inventory depositing in seabed sediments, down to considerable depths. There, the radionuclides are far from being permanently immobilized, since a number of physical and biological processes conspire to bring some of the material back to the water column, in an ongoing biogeochemical cycle.

In the post-Chernobyl years, the geographical distribution of radionuclides in coastal waters was governed by marine processes rather than input processes [86]. Consequently, radionuclide cycling through the oceans and marine foodwebs gained prominence, with more studies on the uptake of radionuclides in aquatic organisms, the derivation of kinetic transfer parameters for use in modeling and, as of recent times, novel mathematical models to represent dynamically the transfer of radionuclides to the biota. The recent Fukushima accident will likely reopen some of these lines of study.

In terms of radiological impact, man-made radionuclides deposited into the marine environment have been found to pose relatively little radiological significance to humans, except in very specific situations. However, the recent emphasis on long-term radiological protection of marine biota has reawakened an interest on the bioaccumulation, potential biomagnification and radiological effects on whole species and ecosystems. The possibility that some species inhabit contaminated locations where humans are not present, and consequently could be more radiologically exposed, merits special attention.

Modeling studies currently underway hold the key for something that up to now has eluded investigators: the development of a meaningful benchmark level for radioactivity in a whole marine ecosystem. Such work may ultimately be able to consider the distortions introduced by radiation on the balance of different species populations, opening new perspectives for the future of radioecology as a science.

Future Directions

Despite an abundance of field research, presently, marine radioecology is singularly lacking in underlying theories and principles that are experimentally testable. Some of the work has an inherent weakness in that it tries to combine data from various sources to derive results, without a clear underlying theoretical basis. It is necessary to improve this situation, studying the behavior of radioactive contaminants in the seas and the radiological impact of substances on man and environment from a “first-principles,” classical science basis. This is particularly important (and necessary) if future reliance on nuclear energy, and all of its potential risks to the marine environment (however small), become the accepted way worldwide.

A direction of particular importance is carrying out laboratory and field studies of radionuclide transfer in marine biota. The concentration factor approach to bioaccumulation (still the most commonly used) is not applicable to dynamic situations, as it does not provide a mechanistic approach to transfer. Biokinetic models can better represent radionuclide transfer to biota in nonequilibrium situations, bringing more realism to predictions of the impact on marine biota arising from pulsed discharges. Such models are required for quantifying the activity retained and how quickly it is cycled and cleared within the marine environment. Much scope for understanding the dynamics of transfer is offered by these modeling and field investigations.

Studies of the speciation of radionuclides in the oceans and the interplay between their different physicochemical forms were initiated in the late 1980s and continue to this date. Much scope is left for such studies in complex and highly dynamic environments such as fjords and estuaries, especially in the vicinity of inputs, which have not been sufficiently investigated.

Subtropical and tropical environments also offer scope for future research, because tropical countries and small islands are among the most active in terms of urbanization and industrialization. Tracking the behavior of heavy metals through their radiotracers offers a way to study threats to such environmentally fragile ecosystems [336, 337].

The long-term fate of the radionuclide inventory in the oceans (especially transuranics) in fine sedimentary deposits is also of interest, and much work remains to be done in this area by providing models and parameters to better understand the stability and final fate of radionuclide inventories.

The current emphasis on investigating the radiological impact of radionuclides on marine ecosystems and risk to nonhuman biota is likely to increase. It is frequently asserted that measures to protect the environment from radiation exposures should focus on the population rather than the individual [338]. An answer to this need is to use population models, and recent studies have begun to focus on the application of such models for the calculation of doses to populations of marine biota in response to chronic radiation-induced changes [338–341]. Population models can be used to study how radiation effects could alter the natural balance of the ecosystem due to the differing radiation sensitivity between predator and prey species in marine food chains [340].

In terms of environmental conservation and sustainability generally, there is a perceived need to integrate protection from ionizing radiation with the approaches adopted for nonradioactive contaminants. Possible integrated assessment approaches are already being outlined, another obvious trend for future research [342].

In the past, radiotracers have helped the scientific community to understand water circulation and marine processes generally. Given the present concerns on climate change, the use of radioactive tracers as indicators of ocean-climate coupling offer potentially a fruitful line of research. The high sensitivity of radionuclide tracers and their chronological ability (using isotope activity ratios) could be used to track changes in thermohaline circulation, which can affect significantly the global climate [343].

Finally, the directions that marine radioecology may take may depend on unforeseen nuclear events. As this article was written, the accident at the Fukushima nuclear power plant in Japan resulted in the dispersal of radionuclides to the Pacific Ocean, in the context of a possible multi-contaminant situation. This may provide an important new direction for radioecology in the future.

Bibliography

Primary Literature

1. Waller EJ (2011) Sources of radiation in the environment including natural radiation, naturally occurring radioactive materials (NORM), technically enhanced materials, weapon tests and nuclear accidents. Encyclopedia of sustainability science and technology. Springer
2. Eisenbud M, Gesell T (1997) Environmental radioactivity – from natural, industrial and military sources, 4th edn. Academic Press, Toronto, Canada
3. Alexakhin RM (2006) Radioecology: history and state-of-the-art at the beginning of the 21(st) century. In: Cigna AA, Druante M (eds) Radiation risk estimates in normal and emergency situations. Springer, Dordrecht, pp 159–168
4. Aarkrog A (1998) A retrospect of anthropogenic radioactivity in the global marine environment. Radiat Prot Dosim 75(1–4):23–31
5. Povinec PP (2003) Worldwide marine radioactivity studies (WOMARS) – Preface. Deep-Sea Res (Part II Top Stud Oceanogr) 50(17–21):2595–2595
6. IAEA (2005) Worldwide marine radioactivity studies (WOMARS) – Radionuclide levels in oceans and seas – Final report of a coordinated research project. In: Povinec P (ed) IAEA-TECDOC-1429, International Atomic Energy Agency, Vienna, 187 pp
7. Lindahl P, Lee S-H, Worsfold P, Keith-Roach M (2010) Plutonium isotopes as tracers for ocean processes: a review. Mar Environ Res 69(2):73–84
8. Solomon KA (1988) Sources of radioactivity in the ocean environment: from low level waste to nuclear powered submarines. J Hazard Mater 18(3):255–262
9. Noshkin VE, Wong KM (1979) Plutonium mobilization from sedimentary sources to solution in the marine environment. Nuclear energy agency seminar on marine radioecology, Tokyo, Japan, 1 Oct 1979
10. Hardy EP, Krey PW, Volchok HL (1973) Global inventory and distribution of fallout plutonium. Nature 2412:444–445
11. Gummer W, Campbell F, Knight G, Richard J (1980) Cosmos 954 – The occurrence and nature of recovered debris. Atomic Energy Control Board AECB- INFO-0006

12. Hardy EP, Krey PW, Volchock HL (1972) Global inventory and distribution of 238Pu from SNAP-9A. United States Atomic Energy Commission Report HASL-250-250 (TID-4500), New York
13. Gascó C (1991) Estudio de la Distribución de Plutonio en el Ecosistema Marino de Palomares Después de una Descarga Accidental de un Aerosol de Transuránidos. PhD Thesis, Universidad Complutense de Madrid, 375 pp
14. Aarkrog A, Dahlgaard H, Nilsson K, Holm E (1984) Studies of plutonium and americium at Thule, Greenland. Health Phys 46:29–44
15. Povinec PP, Gayol J, Togawa O (1999) Global marine radioactivity database (GLOMARD), IAEA-TECDOC-1094, International atomic energy agency, Vienna, pp 481–482
16. Espinosa A, Aragon A, Stradling N, Hodgson A, Birchall A (1998) Assessment of doses to adult members of the public in palomares from inhalation of plutonium and americium. Radiat Prot Dosim 79:161–164
17. Nevissi A, Schell WR (1975) Distribution of plutonium and americium in Bikini Atoll Lagoon. Health Phys 28:539–547
18. Schell WR, Watters RL (1975) Plutonium in aqueous systems. Health Phys 29:589–597
19. Nevissi A, Schell WR (1975) 210Po and 239Pu, 240Pu in biological and water samples from the Bikini and Eniwetak Atolls. Nature 255:321–323
20. Ilus E (2007) The Chernobyl accident and the Baltic Sea. Boreal Environ Res 12:1–10
21. HELCOM (1995) Radioactivity in the Baltic Sea 1984–1991, Helsinki commission and Baltic marine environment protection commission, Baltic sea environment proceedings no. 61, ISSN 0357–2994
22. EC (1994) The radiological exposure of the population of the European community from radioactivity in the Mediterranean Sea. Radiation Protection 70 MARINA-MED Project Report EUR 15564, Luxembourg
23. EC (1990) The radiological exposure of the population of the European community from radioactivity in north European marine waters project MARINA. Radiation protection 47 MARINA-MED project report EUR 12483, Luxembourg
24. Dunster HJ (1998) An historical perspective of radioactive effluents: discharges from Sellafield to the Irish Sea. Radiat Prot Dosim 75(1–4):15–21
25. Kershaw PJ, Woodhead DS, Malcolm SJ, Allington DJ, Lovett MB (1990) A sediment history of Sellafield discharges. J Environ Radioact 12:201–241
26. Leonard KS, McCubbin D, McDonald P, Service M, Bonfield R, Conney S (2004) Accumulation of technetium-99 in the Irish Sea. Sci Total Environ 322(1–3):255–270
27. McCartney M, Kershaw PJ, Woodhead DS, Denoon DC (1994) Artificial radionuclides in the surface sediments of the Irish Sea, 1968–1988. Sci Total Environ 141:103–138
28. Sellafield Ltd (2009) Monitoring our environment. Discharges and monitoring in the United Kingdom. Annual report 2008, Sellafield Ltd., UK, 62 pp
29. Bailly Du Bois P, Boust D, Fievet B, Maro D, Gandon R (1997) Etude de l'impact des rejets de la Hague sur l'environment des mers du nord est de l'Europe
30. Bailly du Bois P, Gueguenait P (1999) Quantitative assessment of dissolved radiotracers in the English Channel: sources, average impact of la Hague reprocessing plant and conservative behaviour (1983, 1986, 1988, 1994). Cont Shelf Res 19:1977–2002
31. Lind OC, Oughton DH, Salbu B, Skipperud L, Sickel MA, Brown JE, Fifield LK, Tims SG (2006) Transport of low 240Pu/239Pu atom ratio plutonium-species in the Ob and Yenisey Rivers to the Kara Sea. Earth Planet Sci Lett 251:33–43
32. Yablokov A (2001) Radioactive waste disposal is seas adjacent to the territory of the Russian federation. Mar Pollut Bull 43(1–6):8–18
33. Calmet DP (1989) Ocean disposal of radioactive waste: status report, in IAEA Bulletin 4/1989. IAEA, Vienna, pp 47–50
34. Salbu B (2001) Actinides associated with particles. In: Kudo A (ed) Plutonium in the environment. Radioactivity in the environment. Elsevier, Oxford, pp 121–138
35. Buesseler KO (1997) The isotopic signature of fallout plutonium in the North Pacific. J Environ Radioact 36:69–83
36. Leonard KS, McCubbin D, Lovett MB (1995) Physico-chemical characterisation of radionuclides discharged from a nuclear establishment. Sci Total Environ 175:9–24
37. Nelson DM, Lovett MB (1978) Oxidation state of plutonium in the Irish Sea. Nature 276:599–602
38. Riglet C, Robouch P, Vitorge P (1989) Standard potentials of the (MO22+/MO2+) and (M4+/M3+) redox systems for neptunium and plutonium. Radiochim Acta 46:85–94
39. Leonard BR (1980) Properties of plutonium isotopes. In: Wick OJ (ed) Plutonium handbook (a guide to technology), vol 1. American Nuclear Society, Illinois, pp 1–8
40. Brown PL (1989) Prediction of formation constants for actinide complexes in solution. Talanta 36(3):351–355
41. Fukai R, Yamato A, Thein M, Bilinski H (1981) Speciation of plutonium in the Mediterranean environment, in techniques for identifying transuranic speciation in aquatic environments. International Atomic Energy Agency, Vienna, pp 37–41
42. Sillén LG, Martell AE (1964) Stability constants of metal-ion complexes. Special publication 17, Chemical Society, London, pp 51–53
43. Rabideau SW, Kline RJ (1958) Kinetics of oxidation-reduction reactions of plutonium. The reaction between plutonium (VI) and plutonium (III) in perchlorate solution. J Phys Chem 62:617–620
44. Rabideau SW (1953) Equilibria and reaction rates in the disproportionation of Pu(IV). J Am Chem Soc 75:798–801
45. Vives i Batlle J (1993) Speciation and bioavailability of plutonium and americium in the Irish Sea and other marine ecosystems. PhD dissertation, Department of Experimental Physics, University College Dublin, Dublin p 347
46. Choppin GR, Bond AH, Hromadka PM (1995) Redox speciation of plutonium. J Radioanal Nucl Chem 219(2):203–210

47. Schulz W (1976) The chemistry of americium. Technical information center, Energy Research and Development Administration, U.S. Department of Commerce, Springfield, 290 pp
48. Pentreath RJ, Woodhead DS, Kershaw PJ, Jefferies DF, Lovett MB (1986) The behaviour of plutonium and americium in the Irish Sea. Rapports et Proces-verbaux des Réunions du Conseil International pour l'Éxploration de la Mer, pp 60–69
49. Vives i Batlle J, Bryan S, McDonald P (2008) A process-based model for the partitioning of soluble, particulate and sediment fractions of plutonium and radiocaesium in the Eastern Irish Sea near Sellafield. J Environ Radioact 99(1):1464–1473
50. Mitchell PI, Vives i Batlle J, Downes AB, Condren OM, Vintro LL, Sanchez Cabeza JA (1995) Recent observations on the physico-chemical speciation of plutonium in the Irish Sea and the Western Mediterranean. Appl Radiat Isotopes 46(11):1175–1190
51. Assinder DJ, Kelly M, Aston SR (1985) Tidal variations in dissolved and particulate phase radionuclide activities in the Esk Estuary, England, and their distribution coefficients and particulate activity fractions. J Environ Radioact 2(1):1–22
52. Nelson DM, Penrose WR, Karttunen JO, Mehlhaff P (1985) Effects of dissolved organic carbon on the adsorption properties of plutonium in natural waters. Environ Sci Technol 79:127–131
53. Nevissi A, Schell WR (1976) Efficiency of a large volume water sampler for some radionuclides in salt and fresh water. In: Cushing CE Jr (ed) Radioecology and energy resources. Dowden, Hutchinson and Ross, Stroudsburg, PA, pp 277–282
54. Nelson DM, Orlandi KA (1985) The role of natural dissolved organic compounds in determining the concentrations of americium in natural waters. In: Bulman RA, Cooper JR (eds) Speciation of fission and activation products in the environment. Elsevier, London, pp 262–268
55. McMahon CA, León Vintró L, Mitchell PI, Dahlgaard H (2000) Oxidation-state distribution of plutonium in surface and subsurface waters at Thule, northwest Greenland. Appl Radiat Isot 52:697–703
56. Lovett MB, Nelson DM (1980) Determination of some oxidation states of plutonium in sea water and associated particulate matter. In: Proceedings of the symposium on techniques for identifying transuranic speciation in aquatic environments. Ispra, Italy. IAEA, Vienna
57. Pentreath RJ, Jefferies DF, Lovett MB, Nelson DM (1980) The behaviour of transuranic and other long-lived radionuclides in the Irish Sea and its relevance to the deep sea disposal of radioactive wastes. In: Marine radioecology, proceedings of 3rd symposium, Tokyo, Japan. OECD, Paris
58. Noshkin VE, Wong KM (1981) Plutonium mobilization from sedimentary sources to solution in the marine environment. In: Proceedings of the marine radioecology, 3rd nuclear energy agency seminar, Paris, 1979, pp 165–178
59. Boust D, Mitchell PI, Garcia K, Condren OM, Leon Vintro L, Leclerc G (1996) A comparative study of the speciation and behaviour of plutonium in the marine environment of two reprocessing plants. Radiochim Acta 74:203–210
60. Jefferies DF, Preston A, Steele AK (1973) Distribution of caesium 137 in British coastal waters. Mar Pollut Bull 4(8):118–122
61. Jefferies DF, Steele AK (1989) Observed and predicted concentrations of caesium-137 in seawater of the Irish Sea 1970–1985. J Environ Radioact 10:173–189
62. Jefferies DF, Steele AK, Preston A (1982) Further studies on the distribution of ^{137}Cs in British coastal waters - I. Irish Sea. Deep-Sea Res Part A 29(6A):713–738
63. Hunt GJ (1985) Timescales for dilution and dispersion of transuranics in the Irish Sea near Sellafield. Sci Total Environ 46:261–278
64. Hunt GJ, Kershaw PJ (1990) Remobilisation of artificial radionuclides from the sediment of the Irish Sea. J Radiol Prot 10(2):147–151
65. Cook GT, MacKenzie AB, McDonald P, Jones SR (1997) Remobilization of Sellafield-derived radionuclides and transport from the north-east Irish Sea. J Environ Radioact 35(3):227–241
66. MacKenzie AB, Cook GT, McDonald P, Jones SR (1998) The influence of mixing timescales and re-dissolution processes on the distribution of radionuclides in northeast Irish Sea sediments. J Environ Radioact 39(1):35–53
67. Bishop GP, Beetham CJ, Cuff YS (1989) Review of literature for chlorine, technetium, iodine and neptunium. Nirex radioactive waste disposal safety studies report NSS/R193, UK Nirex Ltd. Harwell
68. Beasley TM, Lorz HV (1986) A Review of the biological and geochemical behaviour of technetium in the marine environment. J Environ Radioact 3:1–22
69. Rudin MJ, Stanton C, Patterson RG, Garcia RS (1992) National low-level waste management program radionuclide report series. Vol 2: Technetium-99. Idaho national engineering laboratory report, EG and G Idaho Inc., Idaho Falls, 16 pp
70. Schulte EH, Scoppa P (1987) Sources and behaviour of technetium in the environment. Sci Total Environ 64:163–179
71. Sparkes ST, Long SE (1988) The chemical speciation of technetium in the environment - a literature survey, AERE R12743, HMSO, London
72. Wildung RE, McFadden KM, Garland TR (1979) Technetium sources and behaviour in the environment. J Environ Qual 8:156–161
73. Nicholson S, Howorth JM, Sanders TW, Blaine LM (1992) Technetium-99 in the Irish Sea. Department of the environment report no. DOE/HMIP/RR/92.008
74. Leonard KS, McCubbin D, Brown J, Bonfield R, Brooks T (1997) Distribution of technetium-99 in UK coastal waters. Mar Pollut Bull 34(8):628–636
75. McCubbin D, Leonard KS, Brown J, Kershaw PJ, Bonfield RA, Peak T (2002) Further studies of the distribution of technetium-99 and caesium-137 in UK and European coastal waters. Cont Shelf Res 22:1417–1445
76. McCubbin D, Leonard KS, McDonald P, Bonfield R, Boust D (2006) Distribution of Technetium-99 in sub-tidal sediments of the Irish Sea. Cont Shelf Res 26(4):458–473

77. Till JC, Hoffmann O, Dunning DE (1979) A new look at ^{99}Tc releases to the atmosphere. Health Phys 29:695–699
78. Lujanienė G, Šilobritienė B, Jokšas K, Morkūnienė R (2004) Behaviour of radiocesium in marine environment. Environ Res Eng Manage 2(28):23–32
79. Aston SR, Duursma EK (1973) Concentration effects on 137-Cs, 65-Zn, 60-Co and 106-Ru sorption by marine sediments with geological implications. Neth J Sea Res 6(1–2):225–240
80. Ueda S, Hasegawa H, Kondo K (1999) Concentration and speciation of 137Cs in the surface sediments in Brackish Lake Obuchi, Rokkasho village. Radioisotopes 48(11):683–689
81. León Vintró L, Smith KJ, Lucey JA, Mitchell PI (2000) The environmental impact of the Sellafield discharges. In: SCOPE-RADSITE workshop, Brussels, 4–6 Dec 2000, 27 pp
82. Pentreath RJ (1985) General review of literature relevant to water discharges. In: Behaviour of radionuclides released into coastal waters. IAEA-TECDOC-329. International Atomic Energy Agency, Vienna
83. Pentreath RJ, Harvey BR, Lovett MB (1986) Chemical speciation of transuranium nuclides discharged into the marine environment. In: Bulman RA, Cooper JR (eds) Speciation of fission and activation products in the environment. Elsevier, London, pp 312–325
84. Kershaw PJ, Pentreath RJ, Woodhead DS, Hunt GJ (1992) A review of radioactivity in the Irish Sea. A report prepared for the marine pollution monitoring group, Ministry of agriculture, fisheries and food aquatic environment monitoring report 32. MAFF, Directorate of Fisheries Research, Lowestoft, UK, 65 pp
85. Mitchell PI, Condren OM, Vintro LL, McMahon CA (1999) Trends in plutonium, americium and radiocaesium accumulation and long-term bioavailability in the western Irish Sea mud basin. J Environ Radioact 44(2–3):223–251
86. Hirose K (2009) Plutonium in the ocean environment: its distributions and behavior. J Nucl Radiochem Sci 10(1):R7–R16
87. Mitchell PI, Downes AB, León-Vintró L, McMahon CA (2001) Studies of the speciation, colloidal association and remobilisation of plutonium in the marine environment. In: Kudo A (ed) Plutonium in the environment. Elsevier, London, pp 79–104
88. IAEA (2004) Sediment distribution coefficients and concentration factors for biota in the marine environment, Technical reports series no. 422, International Atomic Energy Agency, Vienna
89. IAEA (1985) Sediment Kd's and concentration factors for radionuclides in the marine environment. International Atomic Energy Agency, Vienna
90. Harvey BR, Kershaw PJ (1984) Physico-chemical interactions of long-lived radionuclides in coastal marine sediments and some comparisons with the deep sea environment. In: International symposium on the behaviour of long-lived radionuclides in the marine environment. 1984. Commission of the European Communities, Luxembourg
91. Lansard B, Grenz C, Charmasson S, Schaaff E, Pinazo C (2006) Potential plutonium remobilisation linked to marine sediment resuspension: first estimates based on flume experiments. J Sea Res 55(1):74–85
92. Lepicard S, Raffestin D, Rancillac F (1998) POSEIDON: a dispersion computer code for assessing radiological impacts in a European sea water environment. Radiat Prot Dosim 75(1–4):79–83
93. Simmonds JR, Bexon AP, Lepicard S, Jones AL, Harvey MP, Sihra K, Nielsen SP (2004) Update of the MARINA project on the radiological exposure of the European community from radioactivity in North European marine waters. Annex D: radiological impact on EU member states of radioactivity in north European waters. European Commission report, 190 pp
94. Starik IE (1959) Principles of radiochemistry. Doklady Akademii Nauk SSSR, Translation series AEC-tr-6314, U.S. Atomic Energy Commission
95. Wells ML, Goldberg ED (1991) Occurrence of small colloids in sea water. Nature 353:342–344
96. Wells ML, Goldberg ED (1992) Marine submicron particles. Mar Chem 40:5–18
97. Wells ML, Goldberg ED (1993) Colloid aggregation in seawater. Mar Chem 41:353–358
98. Salbu B (1987) Radioactive tracer techniques in speciation studies. Environ Technol Lett 8:381–392
99. Guo L, Santschi PH, Warnken KW (1995) Dynamics of dissolved organic carbon (DOC) in oceanic environments. Limnol Oceanogr 40:1392–1403
100. Santschi PH, Burd AB, Gaillard J-F, Lazarides A (2005) Transport of materials and chemicals by nanoscale colloids and micro- to macroscale flocs in marine, freshwater, and engineered ecosystems. In: Droppo I, Leppard G, Liss S, Milligan T (eds) Flocculation in natural and engineered environmental systems. CRC Press, Boca Raton, FL
101. Nelson DM, Larsen RP, Penrose WP (1987) Chemical speciation of plutonium in natural waters. In: Pinder JE III, Alberts JJ, McLeod KW, Schreckhise RG (eds) Environmental research on actinide elements. U.S. Department of Energy, Washington, DC, pp 27–48
102. Mudge SM, Hamilton-Taylor J, Kelly M, Bradshaw K (1988) Laboratory studies of the chemical behaviour of plutonium associated with contaminated estuarine sediments. J Environ Radioact 8(3):217–237
103. Penrose WR, Metta DN, Hylko JM, Rinckel LA (1986) The reduction of plutonium (V) by aquatic sediments. J Environ Radioact 5:185–208
104. León Vintró L, Mitchell PI, Smith KJ, Kershaw PJ, Livingston HD (2005) Chapter 3 Transuranium nuclides in the world's oceans. In: Livingston HD (ed) Marine radioactivity. Radioactivity in the environment, vol 6. Elsevier, UK, pp 79–108
105. Nyffeler UP, Li YH, Santschi PH (1984) A kinetic approach to describe trace-element distribution between particles and solution in natural aquatic systems. Geochim Cosmochim Acta 48:1513–1522
106. Oughton DH, Borretzen P, Salbu B, Tronstad E (1997) Mobilisation of ^{137}Cs and ^{90}Sr from sediments: potential sources to Arctic waters. Sci Total Environ 202:155–165
107. Förstner U, Wittmann GTW (1979) Metal pollution in the aquatic environment. Springer, New York, 486 pp

108. Quigley MS, Santschi PH, Guo L, Honeyman BD (2001) Sorption irreversibility and coagulation behavior of 234Th with marine organic matter. Mar Chem 76:27–45
109. Leon Vintro L, Mitchell PI, Condren OM, Downes AB, Papucci C, Delfanti R (1999) Vertical and horizontal fluxes of plutonium and americium in the western Mediterranean and the Strait of Gibraltar. Sci Total Environ 237–238:77–91
110. Noshkin VW, Bowen VT (1973) Concentrations and distributions of long-lived fallout radionuclides in open ocean sediments. In: Radioactive contamination of the marine environment. IAEA, Vienna, pp 671–868
111. Fisher NS, Nolan CV, Fowler SW (1991) Scavenging and retention of metals by zooplankton fecal pellets and marine snow. Deep-Sea Res Part A: Oceanogr Res Pap 38(10):1261–1275
112. Lee B-G, Fisher NS (1992) Decomposition and release of elements from zooplankton debris. Mar Ecol Prog Ser 88:117–128
113. Higgo JJW, Cherry RD, Heyraud M, Fowler SW (1977) Rapid removal of plutonium from the oceanic surface layer by zooplankton faecal pellets. Nature 266:623–624
114. Krishnaswami S, Baskaran M, Fowler SW, Heyraud M (1985) Comparative role of salps and other zooplankton in the cycling and transport of selected elements and natural radionuclides in Mediterranean waters. Biogeochemistry 1(4): 353–360
115. Rodriguez y Baena AM, Fowler SW, Migel JC (2007) Particulate organic carbon: natural radionuclide ratios in zooplankton and their freshly produced fecal pellets from the NW Mediterranean (MedFlux 2005). Limnol Oceanogr 52(3): 966–974
116. Fowler SW, Buat-Menard P, Yokoyama Y, Ballestra S, Holm E, Nguyen HV (1987) Rapid removal of Chernobyl fallout from Mediterranean surface waters by biological activity. Nature 329(6134):56–58
117. Shiomoto A, Hashimoto S, Murakami T (1998) Primary productivity and solar radiation off Sanriku in May 1997. J Oceanogr 54:539–544
118. Alekseev AV, Khrapchenkov FF, Baklanov PJ, Blinov YG, Kachur AN, Medvedeva IA, Minakir PA, Titova GD (2006) Oyashio current – GIWA regional assessment 31: regional definition. Global International Waters Assessments (GIWA) Regional Report, pp 13–21. Available from http://www.unep.org/dewa/giwa/areas/reports/r31/regional_definition_giwa_r31.pdf. Accessed 4 May 2011
119. Terazaki M (1989) Recent large-scale changes in the biomass of the Kuroshio current ecosystem. In: Sherman K, Alexander LM (eds) Biomass yields and geography of large marine ecosystems. AAAS selected symposium 111, Westview, Boulder, pp 37–65
120. Papucci C, Charmasson S, Delfanti R, Gascó C, Mitchell PI, Sánchez Cabeza JA (1996) Chapter 8. Time evolution and levels of man-made radioactivity in the Mediterranean Sea. In: Guegueniat P, Germain P, Metivier H (eds) Radionuclides in the oceans: inputs and inventories. IRSN, France, pp 177–197
121. León Vintró L, Mitchell PI, Smith KJ, Kershaw PJ, Livingston HD (2004) Transuranic radionuclides in the world's oceans. In: Livingston HD (ed) Marine radioactivity, radioactivity in the environment, vol 6. Elsevier, Amsterdam, pp 79–107
122. Fowler SW, Ballestra S, La Rosa J, Fukai R (1983) Vertical transport of particulate-associated plutonium and americium in the upper water column of the northeast pacific. Deep-Sea Res Part A: Oceanogr Res Pap 30(12):1221–1233
123. Bowen VT, Noshkin VE, Livingston HD, Volchok HL (1980) Fallout radionuclides in the Pacific Ocean: vertical and horizontal distributions, largely from GEOSECS stations. Earth Planet Sci Lett 49:411–434
124. Cochran JK, Livingston HD, Hirschberg DJ, Surprenant LD (1987) Natural and anthropogenic radionuclide distributions in the northwest Atlantic Ocean. Earth Planet Sci Lett 84:135–152
125. Nyffeler F, Cigna AA, Dahlgaard H, Livingston HD (1996) - Chapter 1. Radionuclides in the Atlantic Ocean: a survey. In: Radionuclides in the ocean: inputs and inventories. Les Editions de Physique, France
126. Fukai R, Holm E, Ballestra S (1979) A note on vertical distribution of plutonium and americium in the Mediterranean Sea. Oceanol Acta 2(2):129–132
127. Herrmann J, Nies H, Goroncy I (1998) Plutonium in the deep layers of the Norwegian and Greenland Seas. Radiat Prot Dosim 75(1–4):237–245
128. Livingston HD, Kupferman SL, Bowen VT, Moore RM (1984) Vertical profiles of artificial radionuclide concentrations in the Central Arctic Ocean. Geochim Cosmochim Acta 48:2195–2203
129. Smith JN, Ellis KM (1995) Radionuclide tracer profiles at the CESAR Ice Station and Canadian Ice Island in the western Arctic Ocean. Deep-Sea Res 42(6):1149–1470
130. Livingston HD, Povinec PP, Ito T, Togawa O (2001) The behaviour of plutonium in the Pacific Ocean. Radioactiv Environ 1:267–292
131. Ryan TP (2002) Transuranic biokinetic parameters for marine invertebrates–a review. Environ Int 28:83–96
132. NAS (1971) Radioactivity in the marine environment. National Academy of Sciences, Washington, DC, 272 pp
133. McDonald P, Vives i Batlle J, Bousher A, Whittall A, Chambers N (2000) The availability of plutonium and americium in Irish Sea sediments for re-dissolution. Sci Total Environ 267:109–123
134. Clifton J, McDonald P, Plater A, Oldfield F (1997) Relationships between radionuclide content and textural properties in Irish Sea intertidal sediments. Water Air Soil Pollut 99(1–4): 209–216
135. He Q, Walling DE (1996) Interpreting particle size effects in the adsorption of ^{137}Cs and unsupported ^{210}Pb by mineral soils and sediments. J Environ Radioact 30(2):117–137
136. MacKenzie AB, Cook GT, McDonald P (1999) Radionuclide distributions and particle size associations in Irish Sea surface sediments: implications for actinide dispersion. J Environ Radioact 44(2–3):275–296

137. Higgo JJW (1990) Radionuclide Interactions with marine sediments. Nirex radioactive waste disposal safety studies report NSS/R142, UK Nirex Ltd., London
138. Scoppa P, Schulte EH, Secondini A (1982) Chemical form and behaviour of technetium in the marine environment. J Nucl Med Allied Sci 26(3):156
139. Pignolet L, Myttenaere C, Vandercasteele CM, Moreau Z, Cogneau M (1984) Behaviour of technetium in the marine environment: uptake and distribution of Tc into sediments and sedimentary microorganisms. In: Cigna A, Myttenaere C (eds) International symposium on the behaviour of long-lived radionuclides in the marine environment, Proceedings symposium. La Spezia, CEC Report
140. Masson M, Patti F, Colle C, Roucoux P, Grauby A, Saas A (1989) Synopsis of French experimental and in situ research on the terrestrial and marine behaviour of Tc. Health Phys 57(2): 269–279
141. Schreiner F, Fried S, Friedman A (1982) Diffusion of neptunyl (V)- and pertechnetate ions in marine sediments'. In: Topp SV (ed) The scientific basis for nuclear waste management. Elsevier, New York, pp 273–277
142. Burke IT, Livens F, Lloyd JR, Brownd AP, Law GTW, McBeth JM, Ellis BL, Lawson RA, Morris K (2010) The fate of technetium in reduced estuarine sediments: combining direct and indirect analyses. Appl Geochem 25(2):233–241
143. McDonald P, Cook GT, Baxter MS, Thomson JC (1990) Radionuclide transfer from Sellafield to south-west Scotland. J Environ Radioact 12:285–298
144. Kershaw PJ, Denoon DC, Woodhead DS (1999) Observations on the redistribution of plutonium and americium in the Irish Sea Sediments, 1978 to 1996 concentrations and inventories. J Environ Radioact 44:191–221
145. Leonard KS, McCubbin D, Blowers P, Taylor BR (1999) Dissolved plutonium and americium in surface waters of the Irish Sea, 1973–1996. J Environ Radioact 44:129–158
146. Malcolm SJ, Kershaw PJ, Lovett MB, Harvey BR (1990) The interstitial water chemistry of ^{239,240}Pu and ^{241}Am in the sediments of the North-East Irish Sea. Geochim Cosmochim Acta 54:29–35
147. Poole AJ, Denoon DC, Woodhead DS (1997) The distribution and inventory of ^{137}Cs in sub-tidal sediments of the Irish Sea. Radioprotection – Colloques 32(C2):263
148. Leonard KS, McCubbin D, Jenkinson SB, Bonfield RA, McMeekan IT (2008) An assessment of the availability of Tc-99 to marine foodstuffs. Cefas Contract C2170 FOR FSA Project R01062, 81 pp
149. Jones DG, Roberts PD, Strutt MH, Higgo JJ, Davis JR (1999) Distribution of ^{137}Cs and inventories of ^{238}Pu, $^{239/240}$Pu, ^{241}Am and ^{137}Cs in Irish Sea intertidal sediments. J Environ Radioact 44:159–189
150. Jones DG, Kershaw PJ, McMahon CA, Milodowski AE, Murray M, Hunt GJ (2007) Changing patterns of radionuclide distribution in Irish Sea subtidal sediments. J Environ Radioact 96(1–3):63–74
151. Baxter MS, Fowler SW, Povinec PP (1995) Observations on plutonium in the oceans. Appl Radiat Isot 46(11):1213–1223
152. Morse JW, Choppin GR (1986) Laboratory studies of plutonium in marine systems. Mar Chem 20(1):73–89
153. McCubbin D, Leonard KS (1996) Photochemical dissolution of radionuclides from marine sediment. Mar Chem 55(3–4): 399–408
154. Knapińska-Skiba D, Bojanowski R, Radecki Z, Łotocka M (1995) The biological and physico-chemical uptake of radiocesium by particulate matter of natural origin (Baltic Sea). Aquat Ecol 29(3–4):283–290
155. Kim Y, Kim K, Kang H-D, Kim W, Doh S-H, Kim D-H, Kim BK (2007) The accumulation of radiocesium in coarse marine sediment: effects of mineralogy and organic matter. Mar Pollut Bull 54(9):1341–1350
156. Cooper LW, Grebmeier JM, Larsen IL, Dolvin SS, Reed AJ (1998) Inventories and distribution of radiocaesium in arctic marine sediments: influence of biological and physical processes. Chem Ecol 15(1–3):27–46
157. Leonard KS, McCubbin D (2004) Accumulation and remobilisation of 99Tc in the eastern Irish Sea. Environment Report RL 01/04, Centre for Environment, Fisheries and Aquaculture Science, Lowestoft Laboratory, UK, 65 pp
158. McCubbin D, Leonard KS, Emerson HS (1999) The role of thermal and photochemical reactions upon the remobilisation of Pu from an Irish Sea sediment. J Environ Radioact 44:253–273
159. MacKenzie AB, Scott RD, Allan RL, Ben Shaban YA, Cook GT, Pulford ID (1994) Sediment radionuclide profiles: implications for mechanisms of Sellafield waste dispersal in the Irish Sea. J Environ Radioact 23:39–69
160. McCubbin D, Leonard KS (1997) Laboratory studies to investigate short-term oxidation and sorption behaviour of neptunium in artificial and natural seawater solutions. Mar Chem 56(1–2):107–121
161. Aarkrog A, Boelskifte S, Dahlgaard H, Duniec S, Hallstadius L, Holm E, Smith JN (1987) Tc-99 and cesium-134 as long-distance tracers in Arctic waters. Estuar Coast Shelf Sci 24(5):637–647
162. Evans DW, Alberts JJ, Clarke RA (1983) Reversible ion-exchange fixation of caesium-137 leading to mobilization from reservoir sediments. Geochim Cosmocim Acta 47(6):1041–1049
163. Harvey BR (1981) Potential for post depositional migration of neptunium in Irish Sea sediments. In: Impacts of radionuclide releases into the marine environment, IAEA-SM-248/104, IAEA, Vienna, pp 93–103
164. Kershaw PJ, Swift DJ, Pentreath RJ, Lovett MB (1983) Plutonium redistribution by biological-activity in Irish Sea sediments. Nature 306(5945):774–775
165. Officer CB (1982) Mixing, sedimentation rates and age dating for sediment cores. Mar Geol 46:261–278
166. Croker PF (1995) Shallow gas accumulation and migration in the western Irish Sea. In: Croker PF, Shannon PM (eds)

The petroleum geology of Ireland's offshore basins. Geological Society Special Publication No. 93, pp 43–58

167. Pentreath RJ (1984) Alpha-emitting nuclides in the marine environment. Nucl Instrum Methods Phys Res 223:493–501
168. Cournane S, León Vintró L, Mitchell PI (2010) Modelling the reworking effects of bioturbation on the incorporation of radionuclides into the sediment column: implications for the fate of particle-reactive radionuclides in Irish Sea sediments. J Environ Radioact 101:985–991
169. Norton Smith J, Boudreau BP, Noshkin V (1986) Plutonium and 210Pb distributions in northeast Atlantic sediments: subsurface anomalies caused by non-local mixing. Earth Planet Sci Lett 81(1):15–28
170. Nicholson MD, Hunt GJ (1995) Measuring the availability to sediments and biota of radionuclides in wastes discharged to the sea. J Environ Radioact 28:43–56
171. Camplin WC, Durance JA, Jefferies DF (1982) A marine compartment model for collective dose assessment of liquid radioactive effluents. Sizewell enquiry series 4. MAFF Directorate of Fisheries Research, Lowestoft, UK, pp 1–22
172. Gleizon P, McDonald P (2010) Modelling radioactivity in the Irish Sea: from discharge to dose. J Environ Radioact 101:403–413
173. Abril JM, Fraga E (1996) Some physical and chemical features of the variability of Kd distribution coefficients for radionuclides. J Environ Radioact 30(3):253–270
174. Sanchez AL, Gastauda J, Holm E, Roos P (1994) Distribution of plutonium and its oxidation states in Framvaren and Hellvik fjords, Norway. J Environ Radioact 22(3):205–217
175. Pennders RMJ, Prins M, Frissel MJ (1985) The influence of environmental factors on the solubility of Pu, Am and Np in soil-water systems. In: Bulman RA, Cooper JR (eds) Speciation of fission and activation products in the environment. Proceedings CEC-NRPB speciation-85 seminar. Elsevier Applied Science, London, pp 38–46
176. Schell WR, Nevissi A, Huntamer D (1978) Sampling and analysis for Pu and Am in natural waters. Mar Chem 6:143–153
177. Schell WR, Lowman FG, Marshall RP (1980) Geochemistry of transuranium elements at Bikini Atoll. In: Hanson WC (ed) Transuranium elements in the environment. U.S. Department of Energy Publication USDOE/TIC-2280, pp 541–577
178. Assinder DJ, Hamilton-Taylor J, Kelly M, Mudge M, Bradshaw K (1990) Field and laboratory measurements of the rapid remobilisation of plutonium from estuarine sediments. J Radioanal Nucl Chem Art 138:417–424
179. Hamilton-Taylor J, Kelly M, Mudge M, Bradshaw K (1987) Rapid remobilisation of plutonium from estuarine sediments. J Environ Radioact 5(6):409–423
180. Kelly M, Mudge S, Hamilton-Taylor J, Bradshaw K (1988) The behaviour of dissolved plutonium in the Esk Estuary, UK. In: Guary JC, Guéguéniat P, Pentreath RJ (eds) Radionuclides – A tool for oceanography. Elsevier, London/New York, pp 321–330
181. Sholkowitz ER (1983) The geochemistry of plutonium in fresh and marine water environments. Earth Sci review 19:95–161
182. Baskaran M, Ravichandran M, Bianchi TS (1997) Cycling of 7Be and 210Pb in a high DOC, shallow, turbid estuary of southeast Texas. Estuar Coast Shelf Sci 45(2):165–176
183. Lucey JA, Gouzy A, Boust D, Vintro LL, Bowden L, Finegan PP, Kershaw PJ, Mitchell PI (2004) Geochemical fractionation of plutonium in anoxic Irish Sea sediments using an optimised sequential extraction protocol. Appl Radiat Isot 60(2–4):379–385
184. Hamilton EI (1989) Radionuclides and large particles in estuarine sediments. Mar Pollut Bull 20(12):603–607
185. Hamilton EI (1981) Alpha-particle radioactivity of hot particles from the Esk Estuary. Nature 290(5808):690–693
186. Hamilton EI, Williams R, Kershaw PJ (1991) The total alpha particle radioactivity for some components of marine ecosystems. In: Kershaw PJ, Woodhead DS (eds) Radionuclides in the study of marine processes. Elsevier, London, pp 234–244
187. Vives i Batlle J, Wilson RC, McDonald P (2007) Allometric methodology for the calculation of biokinetic parameters for marine biota. Sci Total Environ 388(1–3):256–269
188. Vives i Batlle J, Wilson RC, Watts SJ, Jones SR, McDonald P, Vives-Lynch S (2008) Dynamic model for the assessment of radiological exposure to marine biota. J Environ Radioact 99:1711–1730
189. Brown J, Kolstad AK, Lind B, Rudjord A, Strand P (1998) Technetium-99 contamination in the North Sea and in Norwegian Coastal Areas 1996 and 1997. Strålevern Report 1998: 3. Available at http://www.nrpa.no/dav/07f3957104.pdf. Accessed 7 July 2011
190. Smith V, Ryan RW, Pollard D, Mitchell PI, Ryan TP (1997) Temporal and geographical distributions of 99Tc in inshore waters around Ireland following increased discharges from Sellafield. Radioprotection – Colloques 32:71–77
191. Kasamatsu F, Ishikawa Y (1997) Natural variation of radionuclide 137Cs concentration in marine organisms with special reference to the effect of food habits and trophic level. Mar Ecol Prog Ser 160:109–120
192. Steele AK (1990) Derived concentration factors for cesium-137 in edible species of North Sea fish. Mar Pollut Bull 21(12):591–594
193. Aprosi G, Masson M (1984) Evaluation of experimental studies on technetium transfers to sediments and benthic marine species, and comparison with in situ data. Radioprotection 19(2):89–103
194. Jefferies DF, Hewett CJ (1971) The accumulation and excretion of radioactive caesium by the plaice (Pleuronectes platessa) and the thornback ray (Raia clavata). J Mar Biol Assoc UK 51:411–422
195. Preston A, Jefferies DF (1969) Aquatic aspects in chronic and acute contamination situations. IAEA-SM-117, Report no. STI/PUB/226, IAEA, Vienna, pp 183–211
196. Pentreath RJ, Lovett MB (1978) Transuranic nuclides in plaice (*Pleuronectes platessa*) from the North-Eastern Irish Sea. Mar Biol 48:19–26
197. Smith DL, Knowles JF, Winpenny K (1998) The accumulation, retention, and distribution of ^{95m}Tc in crab (*Cancer pagurus* L.)

and lobster (*Homarus gammarus* L.). A comparative study. J Environ Radioact 40(2):113–135

198. Barnes RD (1974) Invertebrate zoology, 3rd edn. WB Saunders, Philadelphia, 870 pp
199. Busby RG, McCartney M, McDonald P (1997) Technetium-99 concentration factors in Cumbrian Seafood. Radioprotection – Colloques 32(C2):311–316
200. Busby R (1998) The behaviour of technetium-99 in the Irish Sea. PhD thesis, Scottish Universities Research and Reactor Centre, University of Glasgow
201. Knowles JF, Smith DL, Winpenny K (1998) A comparative study of the uptake, clearance and metabolism of techetium in lobster (*Homarus gammarus*) and edible crab (*Cancer pagurus*). Radiat Prot Dosim 75(1–4):125–129
202. Loo LO, Baden SP, Ulmestrand M (1993) Suspension feeding in adult Nephrops norvegicus (L.) and Homarus gammarus (L.) (Decapoda). Neth J Sea Res 31:291–297
203. Pentreath RJ (1981) The biological availability to marine organisms of transuranium and other long-lived nuclides. In: Impacts of radionuclide releases into the marine environment, IAEA STI/PUB/565, Vienna, pp 241–272
204. Swift DJ (1985) The accumulation of ^{95m}Tc from sea water by juvenile lobsters (*Homarus gammarus* L.). J Environ Radioact 2(3):229–243
205. BNFL (1999) Annual report on discharges and monitoring of the environment, environment, health and safety. BNFL, Risley, Warrington, p 152
206. Durand JP, Milcent MC, Goudard F, Paquet F, Germain P, Nafissi T, Pieri J (1994) Chemical behaviour of three radionuclides (cesium, americium and technetium) and their uptake at the cytosolic level in aquatic organisms. Biochem Mol Biol Int 33:521–534
207. BNFL (1998) Annual report on radioactive discharges and monitoring of the environment 1998. British Nuclear Fuel Plc Ltd., UK
208. Miramand P, Germain P, Trilles JP (1989) Histo-autoradiographic localisation of americium (241Am) in tissues of European lobster Homarus gammarus and edible crab Cancer pagurus after uptake from labelled seawater. Mar Ecol Prog Ser 52:217–225
209. Swift DJ (1992) The accumulation of plutonium by the European Lobster (Homarus gammurus L.). J Environ Radioact 16:1–24
210. Ward EE (1966) Uptake of plutonium by the Lobster Homarus vulgaris. Nature 209:625–626
211. Goudard F, Paquet F, Durand J-P, Milcent M-C, Germain P, Pieri J (1994) Biodynamic study of Americium-241 accumulation in the cytosol of the hepatopancreas of the lobster Homarus gammarus. Biochem Mol Biol Int 33:841–851
212. Paquet F, Goudard F, Durand J-P, Germain P, Pieri J (1993) Accumulation of 241Am in subcellular stuctures of the hepatopancreas of the lobster Homarus gammarus. Mar Ecol Prog Ser 95:155–162
213. Conversi A (1985) Uptake and loss of technetium-95m in the crab *Pachygrapsus marmoratus*. J Environ Radioact 2:161–170
214. Swift DJ (1989) The accumulation and retention of Tc-95m by the Edible Winkle (Littorina-Littorea L). J Environ Radioact 9(1):31–52
215. Jackson D, Vives i Batlle J, Parker TG, Whittall AJ, McDonald P (2001) Considerations in optimising the marine discharge regime for ^{99}Tc. In: Proceedings of the British nuclear energy society conference on radiation dose management in the nuclear industry, Windermere, England, 14–16 May 2001
216. Jackson D, Rickard A (1998) The influence of body size and food preparation practices on the uptake and loss of radionuclides in Cumbrian winkles. Radiat Prot Dosim 75(1–4): 155–159
217. McDonald P, Baxter MS, Fowler SW (1993) Distribution of radionuclides in mussels, Winkles and Prawns.1. Study of organisms under environmental-conditions using conventional radio-analytical techniques. J Environ Radioact 18(3):181–202
218. McDonald P, Cook GT, Baxter MS (1992) Natural and anthropogenic radioactivity in coastal regions of the UK. Radiat Prot Dosim 45(1–4):707–710
219. Mitchell J, Vives Batlle J, Ryan TP (1992) Artificial radioactivity in Carlingford Lough. Radiological Protection Institute of Ireland, Dublin
220. Mitchell PI, Vives Batlle J, Ryan TP, McEnri C, Long S, O'Colmain M, Cunningham JD, Caulfield JJ, Larmour RA, Ledgerwood FK (1992) Plutonium, americium and radiocaesium in seawater, sediments and coastal soils in Carlingford Lough. In: Kershaw PJ, Woodhead DS (eds) Radionuclides in the study of marine processes. Elsevier, Amsterdam, pp 265–275
221. Swift DJ (1995) A laboratory study of 239,240Pu, 241Am and 243,244Cm depuration by edible winkles (Littorina Littorea L.) from the Cumbrian Coast (NE Irish Sea) radiolabelled by Sellafield discharges. J Environ Radioact 27(1):13–33
222. Marshall C (1996) Concentration factors for plutonium and americium for mulluscs in the Irish Sea. MSc Dissertation, Environmental Pollution Control Management, Department of Mechanical and Chemical Engineering, Heriot-Watt University, UK
223. Clifton R, Stevens H, Hamilton E (1983) Concentration and depuration of some radionuclides present in a chronically exposed population of mussels (*Mytilus edulis*). Mar Ecol Prog Ser 11:245–256
224. Miramand P, Germain P (1985) Sea water uptake, sediment transfer and histo-autoradiographic study of plutonium (239Pu) and americium (241Am) in the edible cockle Cerastoderma edule. Mar Ecol Prog Ser 22:59–68
225. Dahlgaard H, Nolan C (1999) Long-term rates of radioisotopes of cobalt, zinc, ruthenium, caesium and silver by Mytilus edulis under field conditions. Marine Pollution, IAEA-TECDOC-1094, pp 19–24
226. Guary JC, Fowler SW (1981) Americium-241 and plutonium-237 turnover in mussels (Mytilus galloprovincialis) living in field enclosures. Estuar Coast Shelf Sci 12:193–203

227. Wang WX, Fisher NS, Luoma SN (1996) Kinetic determinations of trace element bioaccumulation in the mussel Mytilus edulis. Mar Ecol Prog Ser 140(1–3):91–113

228. Bonotto S, Robbrecht V, Nuyts G, Cogneau M, Van der Ben D (1988) Uptake of technetium by marine algae autoradiographic localization. Mar Pollut Bull 19(2):61–65

229. Topcuoglu S, Fowler SW (1984) Factors affecting the biokinetics of technetium (Tc-95m) in marine macroalgae. Mar Environ Res 12(1):25–43

230. Gutknecht J (1965) Uptake and retention of ceasium-137 and zinc-65 by seaweeds. Limnol Oceanogr 10(1):58–66

231. Polikarpov GG (1965) Radioecology of aquatic organisms: the accumulation and biological effect of radioactive substances. North Holland, Amsterdam, 314 pp

232. Boisson F, Hutchins DA, Fowler SW, Fisher NS, Teyssie JL (1997) Influence of temperature on the accumulation and retention of 11 radionuclides by the marine alga Fucus vesiculosus (L.). Marine Pollut Bull 35(7–12):313–321

233. Carlson L, Erlandsson B (1991) Effects of salinity on the uptake of radionuclides by fucus vesiculosus L. J Environ Radioact 13:309–322

234. Carlson L, Erlandsson B (1991) Seasonal variation of radionuclides in fucus vesiculosus L. from the Öresund, Southern Sweden. Environ Pollut 73:53–70

235. McDonald P, Cook GT, Baxter MS (1990) A radiological assessment of Scottish edible seaweed consumption. Environ Manage Health 1:17–26

236. Carvalho FP (1985) Experimental studies on biokinetics of americium in benthic marine organisms. In: Bulman JA, Cooper JR (eds) Speciation of fission and activation products in the environment. Elsevier, London, pp 297–315

237. Carvalho FP, Fowler SW (1985) Americium adsorption on the surfaces of macrophytic algae. J Environ Radioact 2:211–317

238. Fowler SW, Benayoun G, Parsi P, Essa MWA (1981) Experimental studies on the bioavailability of technetium in selected marine organisms. In: Proceedings of the international symposium on the impacts of radionuclide releases into the marine environment, Vienna (Austria), 6 Oct 1980. IAEA-SM-248/113

239. Yamamoto T, Fujita T, Ishibashi M (1970) Chemical studies on the seaweeds (25): vanadium and titanium contents in seaweeds. Rec Oceanogr Works Jpn 10:125–135

240. Fisher NS, Nolan CV, Fowler SW (1991) Assimilation of metals in marine copepods and its biogeochemical implications. Mar Ecol Prog Ser 71:37–43

241. Fisher NS, Bjerregaard P, Fowler SW (1983) Interactions of Marine Plankton with transuranic elements.I. Biokinetics of neptunium, plutonium, americium, and californium in phytoplankton. Limnol Oceanogr 28(3):432–447

242. Fisher NS, Rheinfelder JR (1991) Assimilation of selenium in the marine copepod Acartia tonsa studied with a radiotracer method. Mar Ecol Prog Ser 70:157–164

243. Rheinfelder JR, Fisher NS (1991) The assimilation of elements ingested by marine copepods. Science 251:794–797

244. Beasley TM, Gonor JJ, Lorz HV (1982) Technetium - uptake, organ distribution and loss in the mussel, *Mytilus californianus* (Conrad) and the Oyster *Crassostrea gigas* (Thunberg). Mar Environ Res 7(2):103–116

245. Fisher NS (1982) Bioaccumulation of technetium by marine phytoplankton. Environ Sci Technol 16:579–581

246. Warnau M, Fowler SW, Teyssie JL (1996) Biokinetics of selected heavy metals and radionuclides in two marine macrophytes: the seagrass Posidonia oceanica and the alga Caulerpa taxifolia. Mar Environ Res 41(4):343–362

247. Fisher NS, Bjerregaard P, Huynh-Ngoc L (1983) Interactions of marine plankton with transuranic elements. II. Influence of dissolved organic compounds on americium and plutonium accumulation in a diatom. Mar Chem 13:45–56

248. Wilson RC, Watts SJ, Batlle JVI, McDonald P (2009) Laboratory and field studies of polonium and plutonium in marine plankton. J Environ Radioact 100(8):665–669

249. Fisher NS, Bjerregaard P, Fowler SW (1983) Interactions of marine plankton with transuranic elements. III. Biokinetics of americium in euphausiids. Mar Biol 75:261–268

250. Fisher NS, Wente M (1993) The release of trace-elements by dying marine-phytoplankton. Deep-Sea Res (Part I-Topical Stud Oceanogr) 40(4):671–694

251. McCartney M, Rajendran K (1997) 99Tc in the Irish Sea: recent trends. Radioprotection 32:359–364

252. Hamilton EI, Clifton RJ (1980) Concentration and distribution of the transuranium radionuclides Pu-239 + 240, Pu-238 and Am-241 in Mytilus Edulis, Fucus vesiculosus and surface sediment of the Esk estuary. Mar Ecol Prog Ser 3:267–277

253. Nawakowski C, Nicholson MD, Kershaw PJ, Leonard KS (2004) Modelling Tc-99 concentrations in fucus vesiculosus from the north-east Irish sea. J Environ Radioact 77(2): 159–173

254. Ryan TP, Dowdall AM, Long S, Smith V, Pollard D, Cunningham JD (1999) Plutonium and americium in fish, shellfish and seaweed in the Irish environment and their contribution to dose. J Environ Radioact 44(2–3):349–369

255. Thompson N, Cross J, Muller R, Day JP (1982) Alpha and gamma radioactivity in fucus vesiculosus from the Irish sea. Environ Pollut 3:11–19

256. Van der Ben D, Cogneau M, Robbrecht V, Nuyts G, Bossus A, Hurtgen C, Bonotto S (1990) Factors influencing the uptake of technetium by the brown alga fucus-serratus. Mar Pollut Bull 21(2):84–86

257. Sombrito EZ, Banzon RB, de la Mines AS, Bautista E (1982) Uptake of iodine-131 in mussel (*Mytilus smaragdinus*) and algae (*Caulerpa racemosa*). Nucleus J Radioisotope Soc Philippines 22(1):83–89

258. Shunhua C, Qiong S, Xiaokui K (1997) Effects of body size on accumulation and distribution of ^{125}I in the green mussel (*Perna viridis*). China Nuclear Information Centre, Beijing, China

259. Copplestone D, Bielby S, Jones SR, Patton D, Daniel CP, Gize I (2001) Impact assessment of ionising radiation on wildlife. R&D Publication 128, Environment Agency, UK

260. Olsen YS, Vivesi Batlle J (2003) A model for the bioaccumulation of 99Tc in lobsters (*Homarus gammarus*) from the West Cumbrian coast. J Environ Radioact 67(3):219–233
261. Brown J, Børretzen P, Dowdall M, Sazykina T, Kryshev I (2004) The derivation of transfer parameters in the assessment of radiological impacts to arctic marine biota. Arctic 57(3):279–289
262. Fiévet B, Plet D (2003) Estimating biological half-lives of radionuclides in marine compartments from environmental time-series measurements. J Environ Radioact 65:91–107
263. Heling R, Bezhenar R (2009) Modification of the dynamic radionuclide uptake model BURN by salinity driven transfer parameters for the marine foodweb and its integration in POSEIDON-R. Radioprotection 44(5):741–746
264. Lepicard S, Heling R, Maderich V (2004) POSEIDON/RODOS models for radiological assessment of marine environment after accidental releases: application to coastal areas of the Baltic, Black and North Seas. J Environ Radioact 72:153–161
265. Rowan DJ, Rasmussen JB (1996) Measuring the bioenergetic cost of fish activity in situ using a globally dispersed radiotracer (137Cs). Can J Fish Aquat Sci 53:734–745
266. EPIC (2003) The "EPIC" impact assessment framework. Towards the protection of the Arctic Environment from the effects of ionising radiation. In: Brown J, Thorring H, Hosseini A (eds) EPIC (Environmental Protection from Ionising Contaminants) Report ICA2-CT-2000-10032: Østerås, Norway, 175 pp. http: wiki.ceh.ac.uk/display/rpemain/EPIC+reports/. Accessed 7 July 2011
267. Wilson RC, Vives Batlle J, Watts SJ, McDonald P, Parker TG (2007) Uptake and depuration of ^{131}I from labelled diatoms (*Skeletonema costatum*) to the edible periwinkle (*Littorina littorea*). J Environ Radioact 96:75–84
268. Wilson RC, Vives i Batlle J, McDonald P, Parker TG (2005) Uptake and depuration of ^{131}I by the edible periwinkle Littorina littorea: uptake from labelled seaweed (Chondrus crispus). J Environ Radioact 80(3):259–271
269. Vives i Batlle J, Wilson RC, McDonald P, Parker TG (2005) Uptake and depuration of ^{131}I by the edible winkle Littorina littorea: uptake from seawater. J Environ Radioact 78(1):51–67
270. Vives i Batlle J, McDonald P, Parker TG (2002) Studies on the uptake of ^{99}Tc by the edible winkle (*Littornia littorea*). In: Proceedings of the International Conference on Radioactivity in the Environment, Monaco
271. Vives i Batlle J, Wilson RC, McDonald P, Parker TG (2006) A biokinetic model for the uptake and release of radioiodine by the edible periwinkle Littorina littorea. In: Povinec PP, Sanchez-Cabeza JA (eds) Radionuclides in the environment, vol 8. Elsevier, Amsterdam, pp 449–462
272. Schulte EH, Scoppa P, Secondini A (1982) Accumulo e rilascio del tecnezio da parte di alcuni organismi marini: (1) Palaemon elegans. Boll Soc Ital Biol Sper 58:1361–1367
273. West GB, Brown JH, Enquist BJ (1997) A general model for the origin of allometric scaling laws in biology. Science 276(5309):122–126
274. Higley KA, Domotor SL, Antonio EJ (2003) A kinetic-allometric approach to predicting tissue radionuclide concentrations for biota. J Environ Radioact 66(1–2):61–74
275. USDOE (2002) A graded approach for evaluating radiation doses to aquatic and terrestrial biota. Technical Standard DOE-STD-1153-2002. Department of Energy, Washington, DC
276. Carvalho FP, Oliveira JM, Alberto G, Vivesi Batlle J (2010) Allometric relationships of 210Po and 210Pb in mussels and their application to environmental monitoring. Mar Pollut Bull 60(10):1734–1742
277. Kleiber M (1932) Body size and metabolism. Hilgardia 6:315–351
278. Kleiber M (1947) Body size and metabolic rate. Physiol Rev 27(4):511–541
279. Cherry RD, Heyraud M (1991) Polonium-210 and lead-210 in marine organisms: allometric relationships and their significance. In: Kershaw PJ, Woodhead DS (eds) Radionuclides in the study of marine processes. Elsevier, London/New York, pp 309–331
280. Vives i Batlle J, Wilson RC, Watts SJ, McDonald P, Craze A (2009) Derivation of allometric relationships for radionuclides in marine phyla. Radioprotection 44(5):47–52
281. Barnett CL, Kryshev I, Kryshev A, Iospje M, Børretzen P, Golikov V, Shutov V, Kravtsova O, Galeriu D, Vives Lynch SM, Vives i Batlle J, Arkhipov A (2003) Transfer and uptake models for reference arctic organisms. In: Beresford NA, Wright SM, Brown JE, Sazykina T (eds) Environmental protection from ionising contaminants (EPIC) report, European Commission Inco-Copernicus Programme Contract No.: ICA2-CT-2000-10032, May 2003, 75 pp
282. Hendriks AJ (2007) The power of size: a meta-analysis reveals consistency of allometric regressions. Ecol Modell 205(1–2):196–208
283. Lars Føyn L, Heldal HE, Varskog P (2000) Radionuclide uptake and transfer in pelagic food chains of the Barents Sea. In: Brown JE (ed) Radionuclide uptake and transfer in pelagic food chains of the Barents Sea and resulting doses to man and biota, project report under the Transport and Effects Programme. Norwegian Radiation Protection Authority, Oslo, Norway
284. Mathews T, Fisher NS (2008) Evaluating the trophic transfer of cadmium, polonium, and methylmercury in an estuarine food chain. Environ Toxicol Chem 27(5):1093–1101
285. Kryshev I, Sazykina T, Kryshev A, Brown J (2000) Ecological dosimetry models. In: Brown JE (ed) Radionuclide uptake and transfer in pelagic food chains of the Barents Sea and resulting doses to man and biota, project report under the transport and effects programme. Norwegian Radiation Protection Authority, Oslo, Norway
286. Wang WX (2002) Interactions of trace metals and different marine food chains. Mar Ecol Prog Ser 243:295–309
287. Povinec PP, Fowler SW, Baxter MS (1996) Chernobyl and the marine environment: the radiological impact in contex. In: IAEA Bulletin 1/1996, 22 pp

288. Nielsen SP, Keith-Roach M (2004) Update of the MARINA project on the radiological exposure of the European community from radioactivity in North European marine waters. Annex E: critical group exposure. European Commission report, 30 pp
289. Smith JG, Bexon AP, Boyer FHC, Harvey MP, Jones AL, Kindler T, Mercer J, Haywood SM, Verhoef NB, Haverkate BRW, Artmann A (2002) Assessment of the radiological impact on the population of the European Union from European Union nuclear sites between 1987 and 1996. European Commission Radiation protection Report 128, Luxembourg, 38 pp
290. Environment Agency, Food Standards Agency, Northern Ireland Environment Agency and Scottish Environmental Protection Agency (2010) Radioactivity in food and the environment, 2009, RIFE-15 report, 254 pp
291. Jackson D, Lambers B, Gray J (2000) Radiation doses to members of the public near to Sellafield, Cumbria, from liquid discharges 1952–98. J Radiol Prot 20:139–167
292. Hunt GJ, Smith BD (1999) The radiological impact of actinides discharged to the Irish Sea. J Environ Radioact 44(2–3):389–403
293. BNFL (1993–2002) Discharges and monitoring of the environment in the UK (Annual reports – 1993–2002). British Nuclear Fuels Ltd., UK
294. Sellafield Ltd (2010) Monitoring our environment. Discharges and monitoring in the United Kingdom. Annual report 2009, Sellafield Ltd., UK, 65 pp
295. Fegan M, Currivan L, Dowdall A, Hanley O, Hayden E, Kelleher K, Long SC, McKittrick L, Somerville S, Wong J, Pollard D (2010) Radioactivity monitoring of the Irish environment 2008. Report RPII 01/10, Radiological Protection Institute of Ireland, Dublin, 36 pp
296. Smith V, Fegan M, Pollard D, Long S, Hayden E, Ryan TP (2001) Technetium-99 in the Irish marine environment. J Environ Radioact 56:269–284
297. Kershaw PJ, Heldal HE, Mork KA, Rudjord AL (2004) Variability in the supply, distribution and transport of the transient tracer Tc-99 in the NE Atlantic. J Mar Syst 44(1–2):55–81
298. Alexander J, Frøyland L, Hemre G-I, Jacobsen BK, Lund E, Meltzer HM, Skåre JU (2006) Et helhetssyn på fisk og annen sjømat i norsk kosthold. Norwegian Scientific Committee for Food Safety Report – Vitenskapkomiteen for mattrygghet, Oslo, 171 pp
299. Hong GH, Baskaran M, Povinec PP (2004) Artificial radionuclides in the Western North Pacific: a review. In: Shiyomi M, Kawahata H, Koizumi H, Tsuda A, Awaya Y (eds) Global environmental change in the ocean and on land. Terra Scientific Publishing Company, Tokyo, pp 147–172
300. Sazykina TG, Kryshev II (1997) Current and potential doses from Arctic seafood consumption. Sci Total Environ 202(1–3): 57–65
301. NEA (1996) Co-ordinated research and environmental surveillance programme related to sea disposal of radioactive waste. CRESP Final Report 1981–1995, OECD, Paris
302. Pollard D, Long SC, Hayden E, Smith V, Ryan TP, Dowdall A, McGarry A, Cunningham JD (1996) Radioactivity monitoring of the Irish marine environment, 1993 to 1995. Radiological Protection Institute of Ireland, Dublin, 41 pp
303. Livingston HD, Povinec PP (2000) Anthropogenic marine radioactivity. Ocean Coast Manage 43(8–9):689–712
304. Sazykina TG, Kryshev II, Kryshev AI (1998) Doses to marine biota from radioactive waste dumping in the Fjords of Novaya Zemlya. Radiat Prot Dosim 75(1–4):253–256
305. Aarkrog A, Baxter MS, Bettencourt AO, Bojanowski R, Bologa A, Charmasson S, Cunha I, Delfanti R, Duran E, Holm E, Jeffree R, Livingston HD, Mahapanyawong S, Nies H, Osvath I, Pingyu L, Povinec PP, Sanchez A, Smith JN, Swift DJ (1997) A comparison of doses from 137Cs and 210Po in marine food: a major international study. J Environ Radioact 34(1):69–90
306. Togawa O, Povinec PP, Pettersson HBL (1999) Collective dose estimates by the marine food pathway from liquid radioactive wastes dumped in the Sea of Japan. Sci Total Environ 237–238:241–248
307. Aarkrog A (1998) Review of the state of the oceans for radioactivity. In: IAEA 42nd session of the general conference. Scientific forum "Nuclear technology in relation to water resources and the aquatic environment." Available from http://www.iaea.org/About/Policy/GC/GC42/SciProg/gc42-scifor-9.pdf Accessed 22 Mar 2011
308. ICRP (2003) A framework for assessing the impact of ionising radiation on non-human species. ICRP Publication 91, Annals of the ICRP 33(3). Pergamon Press, Oxford
309. IAEA (2005) Proceedings of the IAEA international conference on protection of the environment from the effects of ionising radiation, 6–10 October 2003, Stockholm, Sweden. IAEA CN109/37, International Atomic Energy Agency, Vienna
310. Pentreath RJ, Woodhead DS (2001) A system for protection of the environment from ionizing radiation: selecting reference fauna and flora, and the possible dose models and environmental geometries that could be applied to them. Sci Total Environ 277:33–43
311. ICRP (1991) The 1990 recommendations of the international commission on radiological protection, vol 21, no 1–3. Annals of the ICRP, ICRP 60, Pergamon Press, Oxford
312. Copplestone D, Jones S, Allott R, Merrill P, Vivesi Batlle J (2007) Protection of the Environment from Exposure to Ionising Radiation. In: Shaw G (ed) Radioactivity in the Terrestrial Environment (Radioactivity in the Environment Vol. 10). Pergamon Press, Oxford, 300 pp
313. Watson WS, Sumner DJ, Baker JR, Kennedy S, Reidd R, Robinson I (1999) Radionuclides in seals and porpoises in the coastal waters around UK. Sci Total Environ 234(1–3):1–13
314. Copplestone D, Wood MD, Bielby S, Jones SR, Vivesi Batlle J, Beresford NA (2003) Habitat regulations for stage 3 assessments: radioactive substances authorisations. R&D technical report P3-101/SP1a. Environment Agency, Bristol, 100 pp
315. FASSET (2003). Handbook for assessment of the exposure of biota to ionising radiation from radionuclides in the environment. In: Brown J, Strand P, Hosseini A, Børretzen P (eds)

FASSET deliverable 5 report for the EC 5th framework programme contract FIGE-CT-2000-00102, Norwegian Radiation Protection Authority, Østerås, Norway, 101 pp. https://wiki.ceh.ac.uk/display/rpemain/ERICA+reports. Accessed 7 July 2011

316. Beresford N, Brown J, Copplestone D, Garnier-Laplace J, Howard B, Larsson C-M, Oughton D, Pröhl G, Zinger I (2007) D-ERICA: an integrated approach to the assessment and management of environmental risks from ionising radiation. A deliverable of the ERICA project (FI6R-CT-2004-508847), Swedish Radiation Protection Authority (SSI). Stockholm. Available via http://wiki.ceh.ac.uk/download/attachments/115017395/D-Erica.pdf?version=1. Accessed 4 May 2011
317. Brown JE, Alfonso B, Avila R, Beresford NA, Copplestone D, Pröhl G, Ulanovsky A (2008) The ERICA tool. J Environ Radioact 99(9):1371–1383
318. ICRP (2008) Environmental protection – the concept and use for reference animals and plants for the purposes of environmental protection. In: Valentin J (ed) ICRP publication 108, Annals of the ICRP, Elsevier, Oxford, vol 38 Issue 4–6, 76 pp
319. Kryshev AI, Sazykina TG, Kryshev II, Strand P, Brown JE (2001) Radioecological modelling and the computer codes for calculations of doses to marine biota and man in the Arctic. Environ Modell Softw 16:697–709
320. EC & HC (2003) Canadian environmental protection act 1999. Priority substances list assessment report releases of radionuclides from nuclear racilities (Impact on non-human biota). Environmental Canada, Health Canada Report, 117 pp. Available from: http://www.chemicalsubstanceschimiques.gc.ca/about-apropos/assess-eval/radionuclides-eng.pdf Accessed 1 April 2011
321. Vives i Batlle J, Balonov M, Beaugelin-Seiller K, Beresford NA, Brown J, Cheng J-J, Copplestone D, Doi M, Filistovic V, Golikov V, Horyna J, Hosseini A, Howard BJ, Jones SR, Kamboj S, Kryshev A, Nedveckaite T, Olyslaegers G, Pröhl G, Sazykina T, Ulanovsky A, Vives-Lynch S, Yankovich T, Yu C (2007) Inter-comparison of unweighted absorbed dose rates for non-human biota. Radiat Environ Biophys 46(4):349–373
322. Beresford NA, Barnett CL, Brown J, Cheng JJ, Copplestone D, Filistovic V, Hosseini A, Howard BJ, Jones SR, Kamboj S, Kryshev A, Nedveckaite T, Olyslaegers G, Saxén R, Sazykina T, Vives i Batlle J, Vives-Lynch S, Yankovich T, Yu C (2008) Inter-comparison of models to estimate radionuclide activity concentrations in non-human biota. Radiat Environ Biophys 47(4):491–514
323. Beresford NA, Barnett CL, Brown JE, Cheng JJ, Copplestone D, Gaschak S, Hosseini A, Howard BJ, Kamboj S, Nedveckaite T, Smith JT, Vives I, Batlle J, Vives-Lynch S, Yu C (2010) Predicting the radiation exposure of terrestrial wildlife in the Chernobyl exclusion zone: an international comparison of approaches. J Radiol Prot 30:341–373
324. Brown J, Strand P, Hosseini A, Børretzen P (2003) Handbook for assessment of the exposure of biota to ionising radiation from radionuclides in the environment. FASSET (Framework for assessment of environmental impact) deliverable 5 report, contract no FIGE-CT-2000-00102, 101 pp
325. UNSCEAR (1996) Sources and effects of ionizing radiation. Scientific annex: effects of radiation on the environment, United Nations scientific committee on the effects of atomic radiation report to the general assembly, United Nations, New York, 86 pp
326. Andersson P, Garnier-Laplace J, Beresford N, Copplestone D, Howard B, Howe P, Oughton D, Whitehouse P (2009) Protection of the environment from ionising radiation in a regulatory context (PROTECT): proposed numerical benchmark values. J Environ Radioact 100:1100–1108
327. Woodhead DS (1984) Contamination due to radioactive materials. In: Kinne O (ed) Marine ecology, vol v, part 3: pollution and protection of the seas – radioactive materials, heavy metals and oil. Wiley, New York, NY, pp 1111–1287
328. Vives Batlle J, Wilson RC, Watts SJ, Jones SR, McDonald P, Vives-Lynch S (2008) Dynamic model for the assessment of radiological exposure to marine biota. J Environ Radioact 99(1):1711–1730
329. Sazykina TG, Kryshev II (2002) MARINA II – update of the MARINA project on the radiological exposure of the European community from radioactivity in North European marine waters. Annex F: assessment of the impact of radioactive substances on marine biota of North European waters. EC Radiation Protection Report 132, Directorate-General Environment, Brussels, 44 pp
330. Woodhead DS (1980) Marine disposal of radioactive waste. Helgolander Meeresunters 33:122–137
331. Chambers D, Muller E, Saint-Pierre S, Le Bar S (2005) Assessment of marine biota doses arising from radioactive discharges to the sea by the Cogema La Hague facility – a comprehensive case study. In: Proceedings of the international conference on protection of the environment from the effects of ionizing radiation, Stockholm, 6–10 Oct 2003, pp 159–174
332. OSPAR (2008) Assessment on impact of anthropogenic sources of radioactive substances on marine biota, Radioactive substances series, OSPAR commission report no. 381, 978-1-906840-22
333. Beresford N, Howard B (2005) Application of FASSET framework at case study sites, ERICA deliverable D9. Available from wiki.ceh.ac.uk/display/rpemain/ERICA + reports Accessed 22 March 2011
334. Proehl G, Telleria D, Louvat D (2010) The activities of the IAEA in developing standards on radiological protection of the environment. In: Proceedings of the third European IRPA congress 2010 June 14–16, Helsinki, Finland, 11 pp
335. Livingston HD, Povinec PP (2002) A millennium perspective on the contribution of global. Fallout radionuclides to ocean science. Health Phys 82(5):656–668
336. IAEA (2004) Discovering bio-indicator species for heavy metals in the lagoon of New Caledonia. In: Marine

Environment News, Newsletter of the IAEA marine Environmental Laboratory, Monaco, vol 2, no 1, March 2004

337. Aarkrog A (1994) Past and recent trends in radioecology. Environ Int 20(5):633–643
338. Woodhead DS (2003) A possible approach for the assessment of radiation effects on populations of wild organisms in radionuclide-contaminated environments? J Environ Radioact 66:181–213
339. Vives i Batlle J, Wilson RC, Watts SJ, McDonald P, Jones SR, Vives-Lynch SM, Craze A (2009) Approach to the assessment of risk from chronic radiation to populations of European lobster, *Homarus gammarus* (L.). Radiat Environ Biophys doi:10.1007/s00411-009-0251-y
340. Wilson RC, Vivesi Batlle J, Watts SJ, McDonald P, Jones SR, Vives-Lynch SM, Craze A (2010) Approach to the assessment of risk from chronic radiation to populations of phytoplankton and zooplankton. Radiat Environ Biophys 49:87–95
341. Kryshev AI, Sazykina TG, Sanina KD (2008) Modelling of effects due to chronic exposure of a fish population to ionizing radiation. Radiat Environ Biophys 47(1): 121–129
342. Copplestone D, Howard BJ, Brechignac F (2004) The ecological relevance of current approaches for environmental protection from exposure to ionizing radiation. J Environ Radioact 74:31–41
343. Kershaw PJ (2004) Radiotracers as new barometers of ocean-climate coupling. In: Marine Environment News, Newsletter of the IAEA marine environmental laboratory, Monaco, vol 2, no 1, March 2004

Books and Reviews

Ancellin J, Gueguениat P, Germain P (1979) Radioecologie marine: etude du devenir des radionucleides rejetes en milieu marin et application a la radioprotection. Eyrolles, Paris

Gueguениat P, Germain P, Metivier H (1996) Radionuclides in the oceans: inputs and inventories. Institut de Protection et de Surete Nucleaire, IPSN, France, ISBN 2-86883-285-7

Jones SR (2000) Marine radioecology – five decades of research in the Irish Sea. In: Inaba J, Hisamatsu S, Ohtsuka Y (eds) Proceedings of the international workshop on distribution and speciation of radionuclides in the environment, Rokkasho, Aomori, Japan 11–13 October 2000. Institute for Environmental Sciences, Japan, pp 85–92

Livingston HD (2004) Marine radioactivity. Pergamon Press, Oxford, 310 pp

Matishov DG, Matishov GG (2010) Radioecology in Northern European Seas. Springer, Berlin, 337 pp

Stone R (2002) Radioecology's coming of age – or its last Gasp? Science 297:1800–1801

Sheppard SC (2003) An index of radioecology, what has been important? J Environ Radioact 68:1–10

Whicker FW, Schultz V (1982) Radioecology: nuclear energy and the environment, vol II. CRC Press, Boca Raton, 228 pp

Radiation Assessment, Use of Transfer Parameters

Nicholas A. Beresford
Radioecology Group, Centre for Ecology & Hydrology – CEH Lancaster, Lancaster Environment Centre, Bailrgg, Lancaster, UK

Article Outline

Glossary
Definition of the Subject
Introduction
Application of Transfer Parameters in Human Food Chain Modeling
Radiological Assessment of the Environment
Future Directions
Bibliography

Glossary

Aggregated transfer parameter (T_{ag} m^2 kg^{-1}) Ratio of the activity concentration in an organism (its tissues or milk) and the activity concentration in soil expressed on a per unit area basis (Bq m^{-2}).

Concentration ratio (CR) Ratio of radionuclide activity concentrations in any two compartments. Often used to describe relationship between organism activity concentrations and those in appropriate medium (soil, air, or water) and being recommended as approach to describe relative activity concentrations in animal tissues/milk/egg and those in the animals' diet.

Transfer coefficient (F_f d kg^{-1}F_m d L^{-1}) The ratio of the activity concentration of a radionuclide in the tissues (F_f) or milk (F_m) of an animal to the daily intake of radionuclide in the diet.

Wildlife Wildlife is defined here to include all non-domesticated plants, animals, and other organisms including feral species (i.e., non-native self-sustaining populations).

Definition of the Subject

There had been some consideration of radioactivity from natural sources within the environment prior to World War II. However, the study of the behavior of radionuclides in the environment, driven by the need to assess potential impacts on humans and the environment, has largely evolved since the 1940s following the first atomic bomb tests and subsequent testing programs (predominantly in the 1950s–1960s), and the development of nuclear technologies.

Historically, radiological protection has concentrated on ensuring that humans are protected from the effects of ionizing radiation. Until comparatively recently there has been little regulatory requirement to actively demonstrate the protection of wildlife specifically. In order to assess the contribution of foodstuffs to the internal doses received by humans a number of models have been developed. Many such models rely upon "transfer parameters" which are typically ratios describing the relative activities or activity concentrations between two compartments. Considerable effort has gone into defining such parameters over the last 50–60 years, and compendia of recommended values are published by bodies such as the International Atomic Energy Agency (IAEA).

Starting in the 1990s the anthropogenic focus of radiological protection has been increasingly challenged and there is now legislation in some countries (e.g., UK, USA, Canada, Finland and Sweden) requiring radiological assessment of wildlife. In response to this a number of models/software packages have been developed for application in environmental assessment. These of course require parameterization. The vast numbers of radionuclides and species which may require consideration present challenges to those developing environmental assessment approaches.

Introduction

Exposure of humans and other species to ionizing radioactivity as a result of nuclear power production may occur as a result of planned (authorized) or unplanned releases of radioactive substances from all stages of the nuclear fuel cycle (i.e., from mining to disposal). Other exposure routes include: discharges from facilities that use radioactive substances (e.g., hospitals); industries such as oil and gas extraction and the burning of coal, which release naturally occurring radionuclides to the environment; legacies of nuclear weapons testing and historical accidents; and exposure to radionuclides (such as ^{40}K, ^{238}U-series, ^{232}Th-series, etc.), which are present naturally within the environment.

Methods to assess the risk from these sources of ionizing radiation to humans and other species are required. While exposure to radiation may result from both external and internal sources, this entry considers the estimation of the transfer of radionuclides to human foodstuffs and wildlife (i.e., an essential step in estimating internal exposure). [Readers interested in the methods applied to estimate absorbed dose rates from external and internal sources should consult ICRP [1–4] (for consideration of humans) and Ulanovsky et al. [5] (for consideration of wildlife)]. While ideally sufficient measurements of the radionuclide activity concentrations in human foodstuffs and wildlife would be available to enable robust assessment of internal dose rates, this is not always feasible not least because assessments often need to consider future planned releases. Therefore, models to predict radionuclide concentrations in foodstuffs and wildlife are required. Depending upon their aim such models may need to encompass many processes including: dispersion modeling; interaction with abiotic components of the environment; deposition and interception by vegetation surfaces; surface runoff; and biological uptake and loss. However, many models utilize transfer parameters which are empirical ratios of typically the activities or activity concentrations in different compartments. These can then be used to estimate the radionuclide concentration in a human foodstuff or wild species from a known activity concentration. For instance, the concentration ratio for freshwater fish flesh may be defined as:

$$\mathrm{CR} = \frac{\text{Radionuclide activity concentration in flesh } (\mathrm{Bq\,kg^{-1}} \text{ fresh mass})}{\text{Radionuclide activity concentration in water } (\mathrm{Bq\,L^{-1}})}$$

Therefore, if the activity concentration in water is known then the activity concentration in fish flesh can be estimated as:

$$\mathrm{CR} \times \text{Radionuclide activity concentration in filtered water}$$

As will be evident from the above definition such empirical ratios are a simplification which encompass the many processes described in other entries

(Seaman and Roberts, this section, Rowan this section). The concentration ratio and virtually all of the other empirical ratios discussed below are defined as being at equilibrium. The degree of equilibrium will influence the apparent CR value if measurements are made when equilibrium is not established (e.g., activity concentrations on the tissues of an animal that has been consuming a contaminated diet for a short period will not be equilibrated with those in the diet and hence a low apparent CR value would be derived; plants sampled soon after an atmospheric deposition event will have residual intercepted activity and hence a high apparent CR value would be estimated).

The derivation, application, problems, and advantages of such ratio based approaches are discussed in this entry. The subsequent section presents an overview of transfer parameters for the human food chain focusing on agricultural foodstuffs. The focus of the remainder of the entry is a discussion of the approaches taken to developing models to estimate the transfer of radionuclides to wildlife.

Application of Transfer Parameters in Human Food Chain Modeling

Over the years there have been many compilations of transfer parameters for human foodstuffs which, most notably, include the compilations of Ng and coworkers from Lawrence Livermore Laboratory (USA) in the 1970s and 1980s [6–9] and, more latterly, the IAEA, which has published handbooks of transfer parameters for marine [10], and terrestrial and freshwater environments [11, 12]. The recent update of the terrestrial and freshwater handbook [12] is supported by detailed information on the derivation of the transfer parameter values [13] including extensive publication in the refereed literature (including [14, 15]). The update of the IAEA handbook also prompted the publication of a series of reviews of studies previously only been documented in Russian language publications [16–19] and also consideration of tropical/subtropical ecosystems [20], whereas the previous handbook had only presented data for temperate ecosystems [11].

Soil to Plant Concentration Ratios

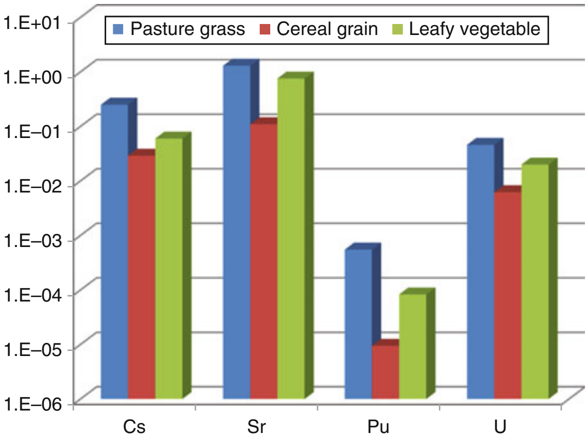
Concentration ratios are the most common approach to estimating the transfer of radionuclides from soil to the edible parts of plants consumed by humans or herbivorous animals. In this instance, CR is most typically defined as:

$$\mathrm{CR} = \frac{\text{Radionuclide activity concentration in plant (Bq kg}^{-1}\text{ dry weight)}}{\text{Radionuclide activity concentration in soil (Bq kg}^{-1}\text{ dry weight)}}$$

Following the recommendations of the International Union of Radioecologists, the IAEA [12] defines CR as being based upon soil activity concentrations determined to a soil depth of 10 cm for grass and 20 cm for all other sample types.

The recent collation of IAEA [12] presents radionuclide CR values for a range of crop types. Where possible, IAEA [12] presents CR values for each crop type by soil textural category (i.e., sand, loam, clay, and organic) in addition to overall summary values by radionuclide and group. Figure 1 presents a comparison of CR values for a selection of the crops and radionuclides from IAEA [12]. Given that CR amalgamates many parameters it is not surprising that for a given radionuclide-crop CR values range over more than two orders of magnitude [13].

For radionuclides with low root uptake rates, measured plant concentrations, and consequently CR values determined in the field, may be dominated by



Radiation Assessment, Use of Transfer Parameters. Figure 1
A comparison of recommended plant CR values for different radionuclides and crops from IAEA [12]

adherent soil. For instance, assuming soil comprises 1% of a vegetation sample on a dry matter basis for a radionuclide with a soil-plant CR of <0.01, the adherent soil will contribute >50% of the activity concentration of the sample; where soil-plant CR values are <0.001 the adherent soil at 1% by dry matter will contribute >90% of the samples activity concentration. Adherent soil can contribute considerably more than 1% by dry matter of pasture grass samples depending upon stocking rate, sward height, and soil type, with values >20% having been recorded [21].

The root uptake of radionuclide generally occurs through soil solution. Interactions of chemical species in soil solution including complexation and precipitation reactions, oxidation-reduction transformations and interactions with soil components (including soil microorganisms) determine the mobility of radionuclides in soils and hence the value of soil-plant CR. The soil-plant CR is therefore influenced by many factors including chemical and physical soil characteristics (percentage clay, percentage organic matter, available stable element concentrations, pH) and land management (e.g., fertilization rates). Soil solution concentrations of Ca and K influence the root uptake of Sr and Cs, respectively. Consequently, lime (calcium carbonate) and potassium fertilizers were successfully used to reduce the uptake of ^{90}Sr and ^{134,137}Cs in agriculturally managed areas affected by the Chernobyl accident [22].

The solid–liquid distribution coefficient (K_d) is used to describe the relative activity concentrations in soil solution and on soil solids, where:

$$K_d(\mathrm{L\ kg^{-1}}) = \frac{\text{Activity concentration in solid phase (Bq kg}^{-1})}{\text{Activity concentration in liquid phase (Bq L}^{-1})}$$

A collation of K_d values can be found in IAEA [12] and a more detailed discussion of the concept in accompanying documentation Vidal et al. [23–26].

Quantifying Radionuclide Transfer to Farm Animals

Since the mid-1960s the most common parameter used to estimate the transfer of radionuclides to farm animals has been the transfer coefficient. This was first proposed by Ward et al. [27] to describe the transfer of radiocesium from the diet to milk of dairy cattle. Ward et al. defined the transfer coefficient (F_m) as:

$$F_m(\mathrm{d\ L^{-1}}) = \frac{^{137}\text{Cs activity concentration in milk (Bq L}^{-1})}{\textit{Daily dietary intake of } ^{137}\textit{Cs (Bq d}^{-1})}$$

The originators reported that the transfer coefficient exhibited less variability between animals within the herd they had studied than expressing transfer as the total amount of radiocesium excreted in milk as a percentage of radiocesium intake. Subsequently, the same workers [28] proposed the transfer coefficient for meat (F_f) as:

$$F_f(\mathrm{d\ kg^{-1}}) = \frac{^{137}\text{Cs activity concentration in meat (Bq kg}^{-1}\text{ fresh weight)}}{\text{Daily dietary intake of } ^{137}\text{Cs (Bq d}^{-1})}$$

The transfer coefficient was adopted by other researchers and modelers as the method of determining transfer for all radionuclides to the milk, meat, and eggs of farm animals; the definition in later publications being the equilibrium ratio of the activity concentration in an animal product to the daily radionuclide intake. In the late 1970s to early 1980s, compendia of recommended transfer coefficient values were being published (e.g., [6–9]) and in 1994, the IAEA included tables of recommended transfer coefficient values in its *Handbook of transfer parameter values for the prediction of radionuclide transfer in temperate environments* [11]. Transfer coefficient values as recommended in such publications were incorporated into many predictive human food chain models (e.g., [29–32]).

Ward and Johnson [33] suggested that this wider use of F_m was justified and that factors such as milk production rate, metabolic rate, soil intake, and stable element intake could be ignored.

In an update to their transfer parameter handbook, IAEA [12] presents recommended transfer coefficient values for: (1) the meat (muscle) of sheep, goats, cattle, pigs, and poultry; (2) the milk of cattle, sheep, and goats; and (3) eggs. In agreement with many other publications the recommended transfer coefficients were generally higher for smaller animals than large animals. This is demonstrated in Figs. 2 and 3, which present a comparison of recommended transfer coefficients for the milk and meat of different animals, respectively, for a few example radionuclides.

For some radionuclides the transfer coefficient for commonly eaten tissues is higher than that to meat. For instance, comparatively high transfer coefficients for liver have been determined for a number of radionuclides such as Pu, Am, Co, and Ag [34–36], while the transfer of Ru to kidney is higher than that to other tissues [35, 36].

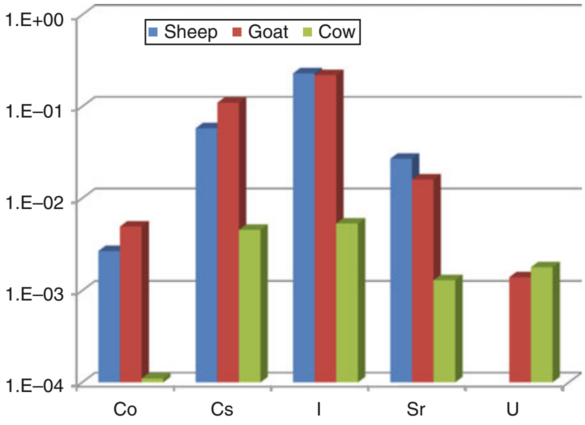
While Ward and Johnson [33] suggested the wide use of transfer coefficients was justified, in Ward and Johnson [37] they also noted that these conclusions were based predominantly on data for the transfer of Cs to cow milk. Subsequently many factors have been demonstrated to influence the transfer coefficient; some examples are described in the following section.

Factors Demonstrated to Influence Transfer Coefficient Values Radiocesium is one of the most studied radionuclides with respect to transfer to farm animals, and factors affecting radiocesium transfer coefficients were the subject of considerable research following the 1986 Chernobyl accident. The dietary source of radiocesium has been demonstrated to have a considerable influence on estimated transfer coefficient values in a number of studies [38–44]. Subsequent research has confirmed that these differences were the consequence of the degree of absorption of radiocesium across the gut [45]. Other studies have demonstrated that the radiocesium transfer coefficient is influenced by lactation [46, 47] and dry matter intake rate [47–49]. However, there is no evidence to suggest that dietary potassium (a chemical analog of cesium) has any significant effect on the transfer of radiocesium to animal-derived food products.

The behavior of radiostrontium in animals has long been known to be determined by that of its chemical analog Ca which is an essential element under homeostatic control (e.g., [50–53]). The absorption of calcium from the diet is inversely proportional to the dietary intake of calcium at a given calcium requirement [54] and an inverse relationship between calcium intake and radiostrontium absorption has been demonstrated in farm ruminants [45]. Howard et al. [55] suggested that the transfer coefficient for radiostrontium to milk (F_{mSr}) could be described by:

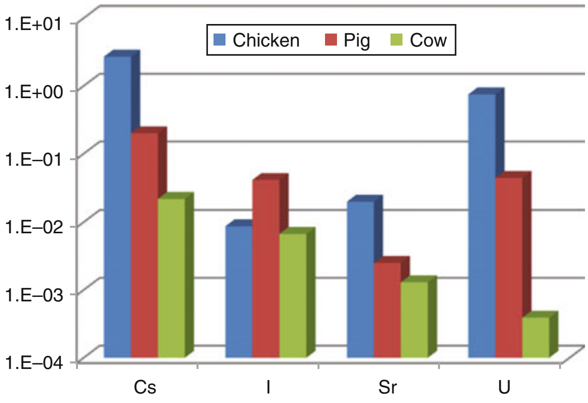
$$F_{mSr}(\mathrm{d\ L^{-1}}) = \frac{OR_{\text{milk-diet}} \times [\mathrm{Ca}]_{\text{milk}}(\mathrm{g\ kg^{-1}})}{\text{Daily dietary intake of Ca }(\mathrm{g\ d^{-1}})}$$

where $OR_{\text{milk-diet}}$ is the observed ratio of the comparative transfers of strontium and calcium from the diet to milk which had previously been shown to be relatively constant with a value of approximately 0.11 [52, 56].



Radiation Assessment, Use of Transfer Parameters. Figure 2

A comparison of recommended milk F_m values for different radionuclides and dairy animals from IAEA [12]



Radiation Assessment, Use of Transfer Parameters. Figure 3

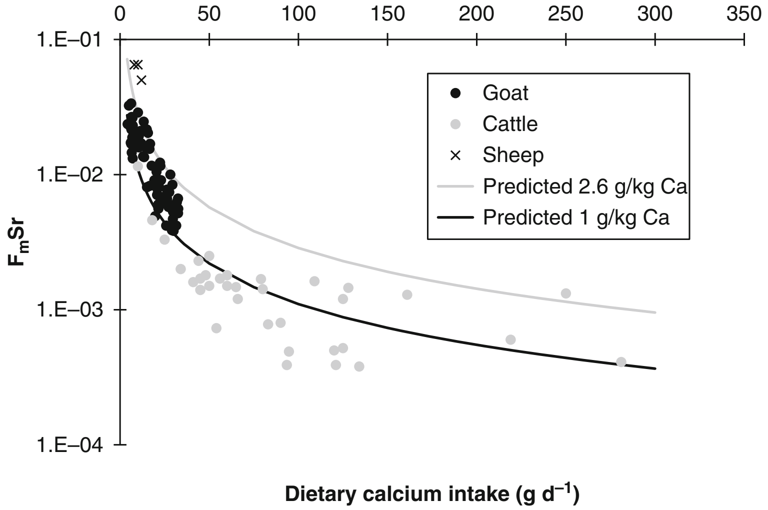
A comparison of recommended meat (muscle) F_f values for different radionuclides and farm animals from IAEA [12]

Assuming a value of calcium in cow milk of 1.15 g kg^{-1}, Howard et al. [55] compared predicted F_{mSr} values with data from a number of published studies and demonstrated a good agreement. Subsequently Beresford et al. [57] demonstrated that the relationship could also be used to predict F_{mSr} for sheep and goat milk sources (assuming appropriate Ca concentrations in the milk of these animals) (Fig. 4). Unlike for radiocesium the dietary source of radiostrontium appears to be relatively unimportant in determining absorption from the gastrointestinal tract and subsequently F_m [45].

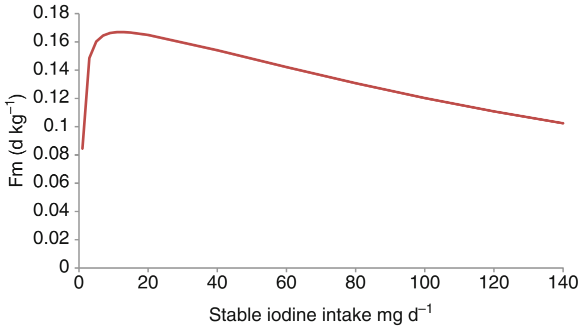
The transfer of radioiodine to milk can be one of the most important routes of exposure to humans in the event of accidental release of radioactivity, for instance, after the 1986 Chernobyl [58, 59] and 1957 Windscale accidents [60]. The absorption of radioiodine from the gastrointestinal tract of ruminant animals is generally complete regardless of stable iodine intake rate of dietary source [45, 61]. The iodine milk transfer coefficient is dependent upon the degree of transfer of absorbed radioiodine from the circulatory system to milk. This appears to be determined by the requirements of the thyroid for iodine [62]. At dietary intake rates below requirement levels low transfer coefficients can be expected compared to animals on diets with optimal levels of iodine [62, 63]. At high rates of iodine intake the transfer of radioiodine to milk increases as the requirements of the thyroid reduce; at extremely high intake rates the transfer of radioiodine to milk is reduced as transfer across the mammary gland saturates [62, 63] (Fig. 5).

The transfer coefficient is defined as the ratio between the activity concentration in an animal-derived food product to the daily radionuclide *intake at equilibrium.* For many radionuclides, activity concentrations in milk reach equilibrium relatively rapidly (e.g., [64, 65]). The uptake rate of some radionuclide to animal tissues is also relatively quick with biological half-lives of a few tens of days [66]. However, the rates of uptake and loss of some radionuclides are comparatively slow such that equilibrium is unlikely to be reached in tissues and even milk within the lifetime of the animal (e.g., [67–69]). For such radionuclides the concept of an “equilibrium” transfer coefficient is therefore questionable (e.g., Pu-isotopes and ^{241}Am).

The transfer of technetium to farm animals has been studied in a number of experimental feeding trials [69–76]. While the main isotope of interest in radiological assessments is the long-lived ^{99}Tc, most



Radiation Assessment, Use of Transfer Parameters. Figure 4
Relationship between calcium intake and F_m for strontium adapted from Beresford et al. [57]; the lines represent predicted values from equation derived by Howard et al. [55] based upon calcium contents in milk of 1 g kg^{-1} (typical for cattle) and 2.6 g kg^{-1} (typical for sheep) respectively



Radiation Assessment, Use of Transfer Parameters. Figure 5

Variation in the radioiodine transfer coefficient for the milk of dairy goats, as predicted by the model of Crout et al. [62]. The stable iodine intakes span suboptimal dietary levels, normal intake rates (1–20 mg day^{-1}), and maximal tolerable intake rates (50 mg kg^{-1} dry matter of feed)

experimental studies have used the short-lived gamma-emitting radioisotopes ^{99m}Tc and ^{95m}Tc. However, the studies of Ennis et al. [69–71] demonstrate variation in the transfer coefficients determined for goat milk for the three technetium radioisotopes. The F_m values of 1.5×10^{-4} d L^{-1} for ^{99m}Tc, 8.5×10^{-4} d L^{-1} for ^{95m}Tc, and 1.1×10^{-2} d L^{-1} for ^{99}Tc were inversely proportional to the specific activity of the three technetium radioisotopes. The authors suggested that this was the consequence of differential rates of reduction of pertechnate to a less available form in the rumen. As a consequence of these isotopes-specific differences in transfer and the lack of data for ^{99}Tc, IAEA [12] does not present recommended transfer coefficient values for technetium [77].

Tritium and ^{14}C are radioactive isotopes of essential macro-elements which constitute the building blocks of animal tissues and feed components; this makes them unusual compared to other radionuclides. Given their potential radiological significance there are relatively few measured transfer coefficients for the two radionuclides. Galeriu et al. [78] presented an approach for the derivation of tritium transfer coefficients based on an understanding of the metabolism of hydrogen in animals and which separately accounted for transfer to and from free (i.e., water) and organically bound tritium. Predictions of the approach compared well to the available data for a range of animal species and products. In Galeriu et al. [79] the approach was expanded to also consider ^{14}C. However, Galeriu et al. cautioned that predicted transfer coefficients for any animal product could not be considered a constant but were dependent upon the animal's metabolic status (see further discussion below). The IAEA [12] does not present transfer coefficient values for ^{3}H and ^{14}C but rather recommend the approach described by Galeriu et al. [78, 79].

Is Transfer Coefficient the Correct Parameter? As already discussed the transfer coefficient has been the accepted parameter to describe the transfer of radionuclides to products of farm animals since the mid-1960s. There has been considerable effort put into understanding the factors which influence transfer coefficients and it is generally accepted that transfer coefficient values for smaller animals are higher than those for larger animals (e.g., see Figs. 2 and 3 above) and those for young animals are higher than those for adults (e.g., [18]). However, the logic of the transfer parameter has increasingly been questioned [9, 36, 47, 77, 80, 81]. In the early 1980s, Ng et al. [9] suggested that size-related differences in estimated transfer coefficients were the consequence of the transfer coefficient incorporating dry matter intake rate which increases with animal size. Consequently, the denominator of the transfer coefficient equation (radionuclide activity concentration in feed × daily dry matter intake rate) will, for instance, be approximately an order of magnitude higher for cattle grazing the same pasture as sheep. However, the elemental composition of the meat and milk of different farm animals does not differ proportionally to size. For instance, Smith et al. [80] considered radiocesium transfer coefficients for the milk of three species and the meat of four species. The milk transfer coefficients they compared ranged from 8×10^{-3} d L^{-1} (cow) to 1×10^{-1} d L^{-1} (goat) while those for meat ranged from 5×10^{-2} d kg^{-1} (beef) to 10 d kg^{-1} (chicken). Rearranging the transfer coefficient equation to derive a food product to dietary concentration ratio (CR), that is:

$$\text{CR} = \text{Dry matter intake rate} \times \text{transfer coefficient}$$

and assuming typical daily dry matter intake rates (taken from [11]), they demonstrated that the CRs for both milk and meat varied by less than twofold.

In a study of the transfer of ^{137}Cs to female sheep throughout a breeding cycle significantly higher F_f values were found for sheep when their lambs had been weaned (F_f *circa* 0.7 d kg^{-1}) than during pregnancy (F_f *circa* 0.3 d kg^{-1}) and lactation (F_f *circa* 0.2 d kg^{-1}) [47]. Documented relationships between dry matter intake rates and protein turnover could credibly explain some of the difference observed in the study. However, when the same data were expressed as concentration ratios the highest value determined was 0.8 d kg^{-1} for lactating sheep with two lambs compared to values in the range 0.3–0.5 for pregnancy and postweaning. This is simply because the dry matter intake rate of the sheep producing milk sufficient to support two lambs was the highest of any experimental group, and therefore this group has the highest denominator in the transfer coefficient equation.

As noted above Galeriu et al. [78, 79] stated that the transfer coefficient for ^{3}H and ^{14}C could not be considered a constant but depends upon metabolic status. For instance, a high yielding dairy animal will ingest more feed and water (and hence C and H) than an animal producing little milk. However, the H and C content of milk and meat will be virtually the same regardless of yield/intake rate, and consequently the transfer coefficient will be lower for high yielding animals. The authors demonstrated that the ratio of the concentration between ^{3}H and ^{14}C activity in meat or milk and the diet is a more robust parameter as it remains relatively constant regardless of the metabolic status of the animal (given that the H and C concentrations in feed and milk/meat remain relatively constant).

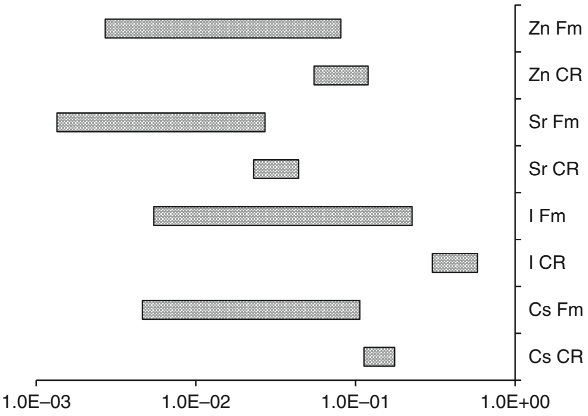
The transfer coefficient can therefore be misleading (e.g., a high transfer coefficient value for one species compared to another does not imply the activity concentrations in this species will be higher). Moreover, as the concentrations of many elements are broadly similar in the milk and meat of different animal species a concentration ratio estimated for one species should give a reasonable estimate of transfer to a species for which there are no data. Howard et al. [81] derived CR values for a number of elements (using radionuclide and stable element data) for: (1) the milk of cattle, goats, sheep, and horses; and (2) the meat of cattle, sheep, and pigs. The resultant CR values were found to differ little between species whereas transfer coefficient values varied across the species by more than an order of magnitude (see Fig. 6 that compares CR and F_m for a number of radionuclides). The CR values estimated by Howard et al. have been included as recommended values in IAEA [12]. However, unfortunately, many radioecological studies do not present data in a manner in which a CR value can be estimated and hence for many radionuclides-animal products, transfer coefficient values still have to be relied upon.

The "Aggregated Transfer Parameter"

Some relatively simple models express transfer to plant- and animal-derived foodstuffs as the aggregated transfer factor (T_{ag}) where:

$$T_{ag}(\mathrm{m^2\ kg^{-1}}) = \frac{\substack{\text{Activity concentration in foodstuff} \\ (\text{Bq kg}^{-1}\text{ dry matter (plant) or fresh weight (animal)})}}{\text{Soil activity concentration (Bq m}^2)}$$

The aggregated transfer parameter is simplification incorporating many processes it tends to be primarily used for seminatural ecosystems including for game, berries and fungi [12]. The T_{ag} is time dependent (e.g., as a consequence of movement down the soil profile) and is often used in association with ecological half-life



Radiation Assessment, Use of Transfer Parameters. Figure 6

A comparison of the range in CR and F_m values [81] for Zn, Sr, I, and Cs to the milk of different animals (ranges include data for cow, sheep, goat, and horse). The figure demonstrates the considerably greater variation in F_m values between species than that in CR values estimates. As with soil-plant CR values there have been some attempts to categorize T_{ag} values on the basis of broad soil type (e.g., Fesenko et al. [82]); Wright et al. [83] related the radiocesium grass T_{ag} to soil organic matter and time after deposition to derive a model to apply in a geographical information system.

Radiological Assessment of the Environment

Although there have been studies to quantify both the exposure of wildlife to ionizing radiation and the potential effects of this exposure (e.g., see Bird this section and Tracy this section), protection of the environment per se has not been a focus of radiation protection until relatively recently. Rather the ICRP [2] statement "the Commission believes that the standards of environmental control needed to protect man to the degree currently thought desirable will ensure that other species are not put at risk" had been relied upon. However, the need to actually demonstrate that the environment is protected is now recognized by a number of international organizations (e.g., [84–87]) and national regulators. In some instances this has been driven by the need to meet conservation legislation [88]. The need for a system to demonstrate environmental protection is recognized in the latest ICRP Recommendations [85].

Approaches to estimate the exposure of wildlife to ionizing radiation have been developed in a number of countries including: the USA [89], England and Wales [90, 91], Korea [92], France [93], Lithuania [94], and Canada [95–97]. Within Europe there have been a series of EC EURATOM funded projects including, most notably, FASSET [98] and subsequently, ERICA [99]. The ICRP have also started to develop a framework for environmental protection based on their reference Animal and Plants concept [100].

An element that all assessment approaches require is a method to predict the transfer of radionuclides to wildlife to be able to determine internal exposure. The most common approaches to doing this are described in this section.

Estimating the Transfer of Radionuclides to Wildlife

All of the current approaches use, at least for some organism, equilibrium ratios between whole-organism activity concentrations and those in environmental media (see [101]), where the whole-organism concentration ratio (CR_{wo}) value is defined for terrestrial ecosystems as:

$$CR_{wo\text{-}soil} = \frac{\text{Activity concentration in biota whole organsim } (\text{Bq kg}^{-1} \text{ fresh weight})}{\text{Activity concentration in soil } (\text{Bq kg}^{-1} \text{ dry weight})}$$

with exceptions, in some models [89, 102], for chronic atmospheric releases of some gaseous radionuclides (such as ^{3}H and ^{14}C), where:

$$CR_{wo\text{-}soil} = \frac{\text{Activity concentration in biota whole organsim } (\text{Bq kg}^{-1} \text{ fresh weight})}{\text{Activity concentration in air } (\text{Bq m}^{-3})}$$

For aquatic ecosystems the majority of approaches calculate CR_{wo} as:

$$CR_{wo\text{-}water} = \frac{\text{Activity concentration in biota whole organsim } (\text{Bq kg}^{-1} \text{ fresh weight})}{\text{Activity concentration in water } (\text{Bq L}^{-1})}$$

If sediment concentrations are known but data for water are lacking, then distribution coefficient (K_d) values are used to estimate concentrations in water. The K_d can also be used to estimate sediment activity concentrations from filtered water concentrations which may be necessary to calculate external dose from sediment. The K_d, which is defined at equilibrium, is determined as:

$$k_d(\text{L kg}^{-1}) = \frac{\text{Activity concentration in sediment } (\text{Bq kg}^{-1} \text{ dry weight})}{\text{Activity concentration in filtered water } (\text{Bq L}^{-1})}$$

The K_d will not be discussed here in detail, however, while models currently source K_d values from existing publications (most especially [10–12]), which summarize data predominantly for application in human assessments. However, most often these consider suspended sediments whereas, for application in assessments of exposure to wildlife it is bed sediments which are of interest. Therefore, the existing reviews may not be ideal for wildlife assessments although

IAEA [103] suggests that the activity concentration of bed sediments is approximately 10% of that of suspended sediments.

There are some models that do not rely solely on organism-media CR_{wo} values but which for some, generally higher, organism types utilize allometric expression relating radionuclide transfer or biological half-life to body mass (e.g., [104–106]) or retention functions (e.g., [106]). Allometric relationships are discussed further in section on "Allometry".

Application of Concentration Ratio in Models and Available Databases

Environment assessment could potentially require the consideration of a large number of organisms. Most of the available environmental radiological assessment approaches have simplified this issue by considering a set of default organisms for which transfer, dosimetric and effects data are collated (e.g., [89, 90, 102]). Some approaches refer to these as "reference organisms" [102, 107, 108]. In selecting these organisms a number of factors have been considered such that they encompass various criteria such as: organisms most likely to be exposed; different tropic levels; organisms sensitive to radioactivity; and protected species. Data are therefore often collated by broad wildlife group such as benthic feeding fish, carnivorous mammal, broadleaf tree, bivalve mollusc, etc.

The ICRP [100] have opted to use a similar approach which they suggest parallels that of the Reference Person [85] approach proposing a set of 12 Reference Animals and Plants (RAPS) (see Table 1) for which effects and transfer, data are being collated and dosimetric parameters estimated. However, the ICRP approach places more emphasis on life stages than the other methodologies considered, with ICRP [100] presenting dose conversion coefficient values for a number of life-stages for some of the RAPs. The RAPs are suggested as "points of reference" for drawing comparisons with sets of information on other organisms. It is acknowledged that the RAPs may not be the direct objects of protection per se and that it may be necessary to establish a "secondary set of Reference Organisms for a specific purpose or geographical area." Since there are no internationally accepted "rules" on classification above Family (or "Super Family") level, the ICRP have suggested that this constitutes the most suitable level of generalization. A RAP is defined as: "a hypothetical entity, with the assumed basic biological characteristics of a particular type of animal or plant, as described to the generality of the taxonomic level of Family, with defined anatomical, physiological, and life-history properties, that can be used for the purposes of relating exposure to dose, and dose to effects, for that type of living organism" [100]. The RAPs are defined in more specific terms than the simplified organisms included in most other approaches (e.g., [91, 103, 109]).

Radiation Assessment, Use of Transfer Parameters. Table 1 ICRP reference animals and plants (RAPs) by three generic habits and the taxonomic family level at which they are defined

RAP	Habitat	Family
Bee	Terrestrial	Apidea
Brown seaweed	Marine	Fucaceae
Crab	Marine	Cancridae
Deer	Terrestrial	Cervidae
Duck	Terrestrial, freshwater	Anatidae
Earthworm	Terrestrial	Lumbricidae
Flatfish	Marine	Pleuronectidae
Frog	Terrestrial, freshwater	Ranidae
Pine tree	Terrestrial	Pinaceae
Rat	Terrestrial	Muridae
Trout	Freshwater, marine	Salmonidae
Wild grass	Terrestrial, freshwater	Poaceae

Most of the current approaches are applied in some form of tier, or iterative, assessment consistent with methods used for other stressors. Such tiered assessments begin with a highly conservative screening tier and progress, if required, to more refined assessments. This is in common with assessment methodologies used for other stressors such as chemicals. In the more comprehensive models the tiered structure is in-built into the software [102, 110]. The different assessment tiers place different requirements on the collation of CR_{wo} values. The highly conservative screening tier typically utilizes maximum or 95th percentile CR_{wo} values. If site-specific data are not available subsequent tiers require best estimate (mean or geometric mean) values, and if the model has probabilistic functionality an associated probability distribution (pdf) (which may also be required to derive 95th percentile values for application in initial screening tiers).

Since 2000 there have been a number of collations of radionuclide transfer data for wild species to support the developing model codes. Subsequent works have generally replaced earlier outputs such that a number of compilations have quickly become out of date. The databases [109, 111] complied for the ERICA Tool (Brown et al. [102]; software freely available from http://www.project.facilia.se/erica/download.html) incorporated those of the earlier European funded FASSET [112] and EPIC [113] projects. Where appropriate they also utilized limited database data compiled for the model (referred to as R&D128]) developed in England and Wales (Copplestone et al. [90, 91]; model spread sheets available from: http://wiki.ceh.ac.uk/x/9wHbBg).

The RESRAD-BIOTA code [110] which implements the USDOE *Graded Approach* [89] contains a default CR database for four generic organisms (terrestrial animal, terrestrial plant, riparian animal, aquatic animal) which was compiled primarily for the initial screen assessment only. Values are generally the maximum CR values (termed B_{iv} in RESRAD-BIOTA) identified by the model developers for any species falling within the appropriate broad organism category (i.e., the value for one radionuclide in a freshwater ecosystem may be for a fish while for another radionuclide the value may be for an invertebrate). For more refined assessments the latest release of RESRAD-BIOTA (code freely available from http://web.ead.anl.gov/resrad/home2/biota.cfm) has some probabilistic functionality and largely recommends the default CR values and associate pdfs from the ERICA Tool. A database of CR values developed to support application of the USDOE Graded Approach can be found online (see http://homer.ornl.gov/nuclearsafety/env/bdac/bcfs.html). However, this is very limited with, in some instances, the denominator not having been specified and its use is not recommended.

Model intercomparison exercises (see further discussion below) have clearly demonstrated that the largest contribution to variability between model predictions, and comparison with available data, is the parameterization of the model transfer components [114–119]. Other studies are in agreement with this conclusion [105, 120]. Based on such observations an IAEA working group recommended that there was a clear need to better share knowledge on the transfer of radionuclides to biota and to provide authoritative collations of those data which are available [101]. Subsequently, the IAEA established a working group to produce a handbook of radionuclide transfer parameters to wildlife (http://www-ns.iaea.org/projects/emras/emras2/working-groups/working-group-five.asp). At the time of writing the handbook is anticipated to be submitted for publication in early 2011 [121]. An online database was established to collate data for the handbook and this will be maintained and updated subsequent to IAEA, the handbook being produced (see https://wiki.ceh.ac.uk/display/rpemain/Wildlife+transfer for links to the database and associated guidance documents). The database has also been utilized by the ICRP to derive CR_{wo} values for their RAPs (a draft report on transfer values for RAPS was released for consultation in July 2010 – see www.ICRP.org). The online database and previous collations (e.g., [109, 111]) utilize stable element data in addition to environmental measurements of radionuclide transfer; typically it is assumed that transfer is element and not isotope dependent.

Data Availability

Current assessment approaches determine whole-organism activity concentrations to subsequently estimate whole-organism dose rates (rather than consideration of target tissues). This approach enables comparison with the available effects data which originate from studies of external gamma exposure (see for instance, Garnier-Laplace et al. [122]). However, many of the available radioecological data for wild species originate from measurements taken for human food chain assessments. These are, therefore, for tissues consumed by humans (most typically muscle). Yankovich et al. [123] provide a series of lookup tables presenting whole-organism: tissue-specific concentration ratios for terrestrial, freshwater, and marine animals which can be used to convert tissue-specific data to whole-organism activity concentrations for the purposes of environmental assessment. While models such as the ERICA Tool [102] nominally estimate whole-organism dose rates for all organism types the geometries used for dosimetric calculations are typically for above ground parts only and CR_{wo} for plants are derived from a variety of above ground plant parts with no attempt to derive a whole-organism activity concentration.

At the time of writing neither the IAEA handbook of wildlife parameters nor the ICRP report presenting CR_{wo} values for their RAPs are finalized and consequently these documents cannot be discussed in any detail. Given this the most comprehensive database of CR_{wo} values for wildlife is that compiled for the ERICA Tool and described by Beresford et al. [112] and Hosseini et al. [113]. The ERICA Tool considers 31 elements and 37 organisms (13 terrestrial organisms, 12 freshwater organisms, and 12 marine organisms); it therefore required 1147 CR_{wo} values. The total number of CR_{wo} values which could be derived from data available in the literature was 453 (i.e., <40% of those required). For some radionuclides and organisms the IAEA handbook [121] and the wildlife transfer database (https://wiki.ceh.ac.uk/display/rpemain/Wildlife+transfer) will considerably expand the amount of data available to those conducting assessments. However, realistically, there will never be a complete set of transfer parameters for all of the potential radionuclide – organism that may be required in assessments. A number of approaches have been derived and used to parameterize existing models [89, 91, 102] to overcome this lack of data. In most cases the resultant parameters are only intended to be used in initial screening tier assessments which aim to be conservative. However, some prospective assessments or the need to consider protected species may require robust extrapolation approaches for application in more refined assessments. As it will present data specifically for species belonging to the taxonomic Family for which each RAP is defined, the amount of data for some RAP-radionuclide combinations will be limited (e.g., there are comparatively large amounts of data available for terrestrial mammals) but very few for species belonging to the Family *Muridae* (i.e., the Reference Rat). Consequently, the ICRP will also need to rely upon extrapolation techniques to provide a complete set of CR_{wo} values.

The next section describes the extrapolation approaches which have been used or suggested for use in radiological assessments of wildlife.

Extrapolation Approaches to Address Lack of Data

A number of extrapolation approaches which have been used to provide values for environmental assessment models are described below. Those having to use such approaches will often have more than one option available to them and will need to make a judgment about which is the most robust based upon the available information they will be applying it to. With the few exceptions which are noted below it is not appropriate here to suggest that any one approach is better than another.

Allometry Allometry, or more properly biological scaling, is the consideration of the effect of size on biological variables. The dependence of a biological variable Y on a body mass M is typically characterized by allometric equations of the form:

$$Y = aM^{b}$$

where a and b are constants.

There are a number of publications summarizing allometric relationships for a wide range of biological variables (e.g., [124]). Many biological phenomena scale as quarter powers of the mass [125, 126], for example: metabolic rates scale as $M^{0.75}$; rates of cellular metabolism and maximal population growth rate, as $M^{-0.25}$; lifespan and embryonic growth and development, as $M^{0.25}$; cross sectional areas of mammalian aortas and tree trunks, as $M^{0.75}$.

Allometric relationships for the biological half-life and dietary transfer coefficient for some radionuclides have been derived [89, 104, 120, 127–129]. While terrestrial mammals and birds have received most attention, Vives i Batlle et al. [104, 129] and Cherry and Heyraud [128] derived allometric relationships between body mass and radionuclide biological half-life and concentration ratio for a number of radionuclides in marine organisms.

In agreement with observations for many other biological parameters the radioecological allometric relationships often scaled to quartile values. Higley et al. [120] and USDOE [89] derived allometric biological half-life relationships for Cs, Sr, I, Co, H, Am, Eu, Pu, Ra, Sb, Tc, Th, U, Zn, and Zr. Of these Cs, Co, Ra, Sb, Sr, U, Zn, and Zr had an exponent of approximately 0.25. Notable exceptions are provided by the actinide elements which scale to about 0.8. Macdonald [127] derived allometric relationship for the tissue-diet transfer coefficient (i.e., F_f as discussed in section on "Application of Transfer Parameters in Human food chain Modeling" for farm animals) for Cs and I for wild animals and birds. The exponent for both elements approached −0.75. For marine organisms Vives i Battle et al. [104] report that (organism-water) CR values for Ru, Ce, Pm/Eu, Ra, Th, Pu, Am, and Cm all scaled allometrically with an exponent of approximately −0.25. However, for marine organisms, while significant, biological half-life values for Tc, Cs, Pu, and Am scaled in the range 0.13–0.2 (Vives i Batlle et al. [104]).

As already noted Macdonald [127] derived allometric relationships describing the transfer coefficient (F_f) of cesium and iodine from feed to the tissues of wild mammals and birds which had similar exponents in the range −0.66 to −0.78. Based on these observations Sheppard [130] hypothesized that the derivation of relationships for other radionuclides required only the estimation of the multiplicand.

Allometric relationships for radionuclide biological half-life [89, 120] were used in the parameterization of the RESRAD-Biota model [110]. The model also allows users to create simple food chain models using these relationships combined with other allometric relationships for dry matter intake and water consumption rates among other parameters. The allometric relationships presented by Higley et al. have been used to provide CR_{wo} values for other models when data are lacking (e.g., [109, 131]).

In the literature discussing allometric relationships in general, there is discussion with regard to both the reasoning for the observation of quarter-power scaling the value of the exponent (e.g., does basal metabolic rate scale to 0.75 or 0.67) (see discussion in Higley and Bytwerk [132]). However, for the purposes of application within radioecological models these nuances are largely of academic interest. Of more relevance in further exploiting the phenomenon is perhaps to understand why some radionuclides and organisms do not scale to approximately quartile values: does this represent a real difference or simply lack of data? Radioecological allometric relationships represent qualitative trends over order of magnitude ranges in body mass and can be used as general predictors across species. They appear fit for this purpose and where predictions using allometric approaches have been compared to available data and/or predictions of other models, they perform adequately [116, 117, 133].

Although some biological traits for plants can be described by allometric functions (e.g., [134, 135]), Higley [136] recently reported that evidence to support the concept of using allometric scaling functions to estimate radionuclide activity concentrations in plants was inconclusive.

Does Size Matter? By algebraic derivation, Beresford et al. [133] suggested that there should be no effect of animal mass on the ratio between the activity concentration in body tissues and feed (for some radionuclides), as the combination of allometric relationships for either biological half-life or F_f with that for dry matter intake (which scales to *circa* 0.75 [137]) removes any mass dependent component from the resulting expression derived for the tissue/organism to diet activity concentration ratio. Considering the allometric relationship for biological half-life, then:

$$\frac{d[\mathrm{WB}]}{dt} = \frac{f_1 \times [\mathrm{Diet}] \times \mathrm{DMI}}{M} - \frac{0.693}{T_{1/2b}} \times [\mathrm{WB}]$$

where [WB] is the wholebody activity concentration of a radionuclide in an animal (Bq kg^{-1} fresh weight), t is time (d), $T_{1/2b}$ is the radionuclide biological half-life (d), DMI is the daily dry matter intake of feed (kg day^{-1}), f_1 is the fraction of ingested radionuclide absorbed from the gastrointestinal tract (dimensionless), and M is body mass (kg). At equilibrium and substituting DMI and $T_{1/2b}$ for allometric relationships assuming scaling factors of 0.75 and 0.25, respectively, this can be rearranged to:

$$\frac{0.693 \times [\mathrm{WB}]}{a_1 M^{0.25}} = \frac{[\mathrm{Diet}] \times a_2 M^{0.75}}{M}$$

which can be rearranged to give a wholebody to dietary activity concentration ratio:

$$\frac{[\mathrm{WB}]}{[\mathrm{Diet}]} = \frac{a_1 M^{0.25} \times a_2 M^{0.75}}{0.693 M} = \frac{a_1 a_2}{0.693}$$

where a_1 and a_2 are the multiplicands of the allometric expressions for $T_{1/2b}$ and DMI, respectively.

Similarly starting from the equation for F_f presented in section on "Application of Transfer Parameters in Human Food Chain Modeling" and the allometric relationship suggested assuming a scaling factor of −0.75:

$$F_f = \frac{[WB]}{a_1 M^{0.25} \times [Diet]} = a_3 M^{-0.75}$$

can be rearranged to give a wholebody to dietary activity concentration ratio:

$$\frac{[WB]}{[Diet]} = a_1 a_2$$

where a_3 the multiplicand of the allometric expressions for F_f. Consequently, while F_f and biological half-life for radionuclides may vary considerably, concentration is relatively independent of size [133, 136] as demonstrated in Table 2. This is in agreement with the suggestion made in section on "Application of Transfer Parameters in Human Food Chain Modeling" that the dietary CR is a more robust parameter for estimating transfer to farm animals than the transfer coefficient. However, as should be apparent from Table 2, while the equilibrium activity concentration is relatively constant across a wide range of body masses the time taken to reach that concentration will increase with size.

The assumption of a constant CR value between wholebody and dietary activity concentrations was used to provide default CR_{wo} values for some terrestrial vertebrates in the ERICA Tool [109].

Radiation Assessment, Use of Transfer Parameters. Table 2 Allometric predictions of radiocesium biological half-life, F_f, and equilibrium activity concentration for organisms of increasing mass

Body mass (kg)	Biological half-life (d)	F_f (d kg^{-1})	Bq kg^{-1}
0.001	2.6	1,400	0.46
0.01	4.4	260	0.47
0.1	7.7	48	0.48
1.0	13.2	8.9	0.49
10	22.8	1.7	0.50
100	39.8	0.31	0.51
1,000	68.0	0.06	0.52

Assume Value for a Similar Organism Where data are lacking for an organism-radionuclide combination the CR_{wo} value for a "similar" organism may be assumed. This approach was used in the ERICA model [109, 111] and Copplestone et al. [91]. For instance, if there are no data for ^{241}Am transfer to birds then a model may assume the CR_{wo} for mammals may be applied (both being warm-blooded vertebrates with a similar trophic levels).

Such decisions are obviously subjective but they can perhaps be given some scientific grounding. Differences in CR values between vascular plant species have been shown to be related to their evolutionary history, or phylogeny [138]. Broadley et al. [139] demonstrated that variation in the accumulation of radiologically relevant metals (Ni, Pb, Zn, Cd, Cr, and Cu) occurred at classification of the order or above. Willey [138] demonstrated significant phylogenetic signals for soil-plant transfer for Cs, Sr, Co, Cl, and Ru with monocots being found to have lower transfer than eudicots. A similar finding was reported in Willey et al. [140] for ^{99}Tc.

For marine organisms Jeffree et al. [141] demonstrated that the rates of uptake of nine radionuclides from water by three species of *chondrichthyans* differed to that by one species of *pleuronectiform* and two species of *perciform teleost* species.

Put simply, a significant phylogenetic signal demonstrates that transfer of the element is likely to be more similar for organisms which are close to each other evolutionary than those which are distant. Willey [138] presents the statistical coding to perform phylogenetic analyses. However, the analysis is rather data hungry.

Based upon an analysis of tree and crop data, Tagami and Uchida [142] suggest that CR values derived for crops (which are relatively numerous) could be applied to predict radionuclide transfer to trees if data for the latter are lacking.

Analog Elements Some approaches have assumed that where CR data are lacking an appropriate value for an element with similar biogeochemical behavior can be assumed. This generally entails consideration of placement on the periodic table and knowledge that the elements behave in a similar manner. For instance, to provide a complete set of CR_{wo} values for the ERICA Tool assumptions such as applying Am values for Pu, Ce values for La, U values for Th, etc., were made if data were not available for a given element [109, 111].

Higley [136] suggests that soil-plant transfer could better be predicted using ionic potential citing work by Railsback [143] and Tyler [144]. Ionic potential is the ratio of ionic charge of a cation or anion to its ionic radius (see http://www.gly.uga.edu/railsback/PT.html). Considering 50 cationic elements Tyler [144] demonstrated a strong correlation ($R^2 = 0.6$) between ionic potential and CR for roots.

However, it is likely that in most instances selecting analog elements based upon period table grouping/prior knowledge of comparative behaviors or ionic potential would result in a similar element being selected.

Specific Activity Models As for human food chain transfer models (see section on "Quantifying Radionuclide Transfer to Farm Animals") conventional modeling techniques used for the majority of elements may not be appropriate to determining the transfer of ^3H and ^{14}C to wildlife. Some approaches [90, 102, 113] therefore use specific activity models which given these two radionuclides are primarily present as reversible gases ($^{14}CO_2$ and ^3HHO) relate activity concentrations in wildlife to those in air (Bq m^{-3}) for terrestrial systems. The derivation of the ^3H and ^{14}C approach adopted in the ERICA Tool is described in Galeriu et al. [145]. Currently, the approach taken by the ERICA Tool for aquatic ecosystems is inconsistent with that taken for terrestrial ecosystems in that empirical CR_{wo} values are used.

"Extreme" Assumptions In order to provide complete CR_{wo} datasets some approaches have used assumptions which could be considered as rather extreme. These include using the kd value for aquatic organisms if no CR_{wo} values are available [91]; applying the maximum available CR_{wo} for the given element regardless of organism [91, 109, 111]; and using a CR_{wo} for the organism-radionuclide from a different ecosystem [111]. Readers are reminded that such assumptions are used to enable initial screening level assessments to be conducted and are intended to yield conservative CR_{wo} values; such default values should not be used in more realistic assessments.

Problems with Providing "Complete" Sets of Cr_{wo} Value As noted in the previous sections, some radiological environmental assessment models in an attempt to enable users to be able to conduct initial screening tier assessments have used various extrapolation techniques to derive CR_{wo} values where data are unavailable. In some cases default values derived by such models are clearly identified in the model and/or accompanying documentation [91, 102].

However, there is a potential problem in providing complete sets of CR_{wo} values in that users may ignore caveats on their application. There is some published evidence that this is occurring [119, 146].

The derivation of transfer parameters where data are lacking is not restricted to assessments of wildlife exposure; IAEA [10] derives values for some radionuclide-organisms using similar approaches for application in assessments of human exposure via seafood consumption. And IAEA [12] presents a discussion of the application of analog elements and species for application in terrestrial and freshwater foodstuff models.

How are Environmental Assessment Models Performing?

The IAEA initiated a working group under its Environmental Modeling for Radiation Safety (EMRAS) program in 2004 to compare and validate models being used and developed to demonstrate protection of the environment (see IAEA [101]) These activities are currently continuing under the IAEA EMRAS II program (see http://www-ns.iaea.org/projects/emras/emras2/working-groups/working-group-four.asp).

The working groups have conducted (and continue to conduct) a number of model-model and model-data intercomparisons. Two of these exercises considered the estimation of dose under assumed conditions of unit activity concentration in medium or organism [114, 115]. Together with the other exercises these demonstrated that for consideration of simplified organism-medium geometries, the differences in dosimetry between the various models add little to the overall uncertainty in radiological assessments of wildlife. However, comparisons of predicted activity concentrations assuming 1 Bq per unit media showed variability in predictions which for some organism-radionuclide combinations varied by more than three orders of magnitude [116]. Predictions were often the most variable for poorly studied organisms (e.g., amphibians, ducks, aquatic mammals). Some of the most extreme variability could be attributed to some users applying default model values derived using the extrapolation approaches described above. In some models the application of such techniques to provide default values aim to be conservative and in most instances comparatively high (i.e., conservative) predictions were made.

Yankovich et al. [119] reports the EMRAS working groups predictions of ^{60}Co, ^{90}Sr, ^{137}Cs, and ^{3}H activity concentrations in a variety of wildlife in Perch Lake (Ontario, Canada). Key findings were:

- Models parameterized using CR_{wo} values derived from laboratory experiments tended to under predict
- Models accounting for water calcium concentrations (see example below) better predicted ^{90}Sr activity concentrations in fish than models applying generic CR_{wo} values
- Predictions were generally poor for aquatic mammals and herpetofauna

An example of the relationships between CR_{wo} and water calcium concentrations used is that of Kryshev [147]:

$$CR_{wo}Sr = \frac{3940}{[Ca^{2+}]}$$

where $[Ca^{2+}]$ is calcium concentration in water (mg L^{-1}).

A similar intercomparison was made for terrestrial organisms in the Chernobyl Exclusion Zone with predictions being made for ^{137}Cs, ^{90}Sr, Pu-isotopes, and ^{241}Am [117]. In many instances predictions were found to be within an order of magnitude of the observed data, however, some predictions were under- or over-estimates by more than two orders of magnitude. Two models applied in the exercise applied allometric biological half-life relationships (see "Allometry"); overall these models performed similarly to those using simple CR approaches.

It is envisaged that the majority of assessed sites will require only a screening assessment. An initial screening levels assessment is designed to be simple requiring minimal inputs and provide conservative results, with the aim of being able to identify sites of negligible concern and to remove them from further consideration with a high degree of confidence. As an example within England and Wales, 600 out of approximately 700 sites with authorizations to discharge radioactivity were identified as not requiring more detailed assessment (i.e., they were "screened" out) [148]. Beresford et al. [118] compared the screening level application of the R&D128 model, RESRAD-BIOTA and the ERICA Tool models, which are available for use by any user (see section on "Application of Concentration Ratio in Models and Available Databases" for links to download software) to data for seven sites. The outputs of the three models varied considerably for a given assessment with different conclusions being inferred with regard to the requirement if the site required any more detailed assessment. A number of factors were identified as contributing to this variability including: CR_{wo} and K_d values, input options, secular equilibrium assumptions, and geometry and exposure scenario. Such wide variation in screening application results gives some concern as to if they can be used with a high degree of confidence.

Future Directions

Many states currently face a number of radiological issues associated with: (1) long-term waste storage; (2) likely expansion of nuclear power productions; (3) legacy sites (most often associated with mining and mineral processing operations); (4) emergency planning (including potential malevolent use of radioactive sources); and (5) adapting to new recommendations and regulation (e.g., recent developments to explicitly demonstrate that the environment is protected). Recent reviews, such as IAEA [12], demonstrate that knowledge of the transfer of some radionuclides, including those identified as of concern with regard to some of these issues, is poor.

The majority of the parameters discussed above are relatively simplistic equilibrium ratios (although these may often be used within models having some dynamic parameters). These offer a pragmatic approach to predicting radionuclide concentrations in human foodstuffs and wildlife. However, they often amalgamate many biological-chemical-physical processes and hence have a high degree of associated uncertainty.

In some instance these simplistic approaches have been refined and uncertainty reduced by categorizing transfer parameters on the basis of environmental variables (e.g., the categorization of soil-plant CR in IAEA [12] on the basis of soil type). Derivation of relationships between the transfer of Sr to milk and the dietary calcium intake represents a relatively simple approach to making best use of scientific knowledge (see section on "Factors Demonstrated to Influence Transfer Coefficient Values"). Similarly, CR values for freshwater fish have been related to water Ca and K concentrations for Sr and Cs, respectively [149, 150].

An essential first step in defining future research requirements is to assess where current models are fit for purpose. It is highly unlikely that detailed mechanistic knowledge will be gained for all radionuclides which require assessment. However, is such knowledge required for all radionuclides or is it probable that some which have to be assessed (as they constitute authorized discharges) are unlikely to contribute significantly to dose rates? Similarly, it may be that an equilibrium CR_{wo} is adequate for use in assessments of exposure of wildlife for the long time frames and exposure routes considered in assessments for waste repositories. However, some element of dynamic modeling may be required when considering chronic atmospheric release as a consequence of interception on vegetation surfaces [151].

The application of foodstuff, or organism, to medium concentration ratios will always have associated uncertainty. Depending upon the purposes of the assessment, or radionuclide and exposure pathway considered, such uncertainty may be acceptable. However, there will be circumstances where there is a requirement to better predict/understand radionuclide behavior. The development of models which utilize readily available/measurable environmental parameters should be encouraged as these will lead to models which can be applied globally. A good example of such a model for terrestrial systems is that described by Gillett et al. [152] which uses soil parameters such as exchangeable K concentration, pH, percentage clay, and percentage organic matter to predict the uptake of radiocesium by vegetation. For other radionuclides it is likely that environment chemistry models developed for stable elements (e.g., [153]) could provide a basis for the development of radioecological models.

As resources will always be limited well founded extrapolation approaches should be pursued. These will be especially useful to provide parameters for models predicting transfer to wildlife although they also have application in human food chain models (e.g., dietary CR represents a more generic approach than does the generally accepted transfer coefficient) (see section on "Is Transfer Coefficient the Correct Parameter?").

Sensitive analytical techniques such as ICP-MS, ICRP-OES, and neutron activation offer the possibility of obtaining data for a large number of elements of relevance to radiological assessments from the same sample and such approaches are increasingly being used (e.g., Higley [136], Takata et al. [154]; Tagami and Uchida [142]). Such techniques offer a cost- and resource-effective approach to obtaining data for poorly studied radionuclides. However, there is perhaps a requirement to more fully assess any limitations to the application of stable element data in models of radionuclide behavior (Uchida et al. [155] considers this for transfer to rice).

Where possible research to improve human food chain and wildlife assessment models should be integrated as the key processes determining transfer will often be the same (e.g., water chemistry, soil characteristics).

The potential implications of climate change on radiological assessments have to date received little consideration. While there may be some circumstances where climate directly impacts on radionuclide transfer (e.g., period of soil freezing in Arctic regions; ^{3}H transfer), largely the impact of climate is likely to be indirect (e.g., influence of climate on soil, land use, crop types, farm animals, and their management). However, the existing data are biased toward northern temperate ecosystems.

This entry focused on the "transfer parameters" used in radiological assessments. Other parameters (weathering rates, interception factors, resuspension rates, gastrointestinal absorption coefficients, dispersion coefficient, biological half-lives, etc.) are used in models and an overview of many of these can be found in IAEA [13].

Bibliography

Primary Literature

1. International Commission on Radiological Protection (ICRP) (1977) Recommendations of the international commission on radiological protection, ICRP publication 26. Pergamon, Oxford
2. International Commission on Radiological Protection (ICRP) (1991) Recommendations of the international commission on radiological protection, ICRP publication 60; Annals of the ICRP 21. Pergamon, Oxford
3. International Commission on Radiological Protection (ICRP) (1990) Age-dependent doses to members of the public from intake of radionuclides: Part 1, vol 20/2, ICRP publication 56; Annals of the ICRPl. Pergamon, Oxford
4. International Commission on Radiological Protection (ICRP) (1996) Age-dependent doses to members of the public from intake of radionuclides: Part 5 compilation of ingestion and inhalation dose coefficients, vol 26/1, ICRP publication 72; Annals of the ICRPl. Pergamon, Oxford
5. Ulanovsky A, Pröhl G, Gomez-Ros JM (2008) Methods for calculating dose conversion coefficients for terrestrial and aquatic biota. J Environ Radioactiv 99:1440–1448
6. Ng YC (1982) A review of transfer factors for assessing the dose from radionuclides in agricultural products. Nucl Safety 23:57–71
7. Ng YC, Colsher CS, Quinn DL, Thompson SE (1977) Transfer coefficients for the prediction of the dose to man via the forage-cow-milk pathway from radionuclides released to the biosphere, UCRL-51939. Lawrence Livermore Laboratory/University of California, Livermore/California, 94550
8. Ng YC, Colsher CS, Thompson SE (1979) Transfer factors for assessing the dose from radionuclides in agricultural products. In: Proceedings of an international symposium on the biological implications of radionuclides released from nuclear industries, IAEA-SM-237/54. International Atomic Energy Agency, Vienna
9. Ng YC, Colsher CS, Thompson SE (1982) Transfer coefficients for assessing the dose from radionuclides in meat and eggs: final report, NUREG/CR-2976. US Nuclear Regulatory Commission, Washington, DC
10. International Atomic Energy Agency (IAEA) (2004) Sediment distribution coefficients and concentration factors for biota in the marine environment, Technical Reports Series No. 422. International Atomic Energy Agency, Vienna
11. International Atomic Energy Agency (IAEA) (1994) Handbook of parameter values for the prediction of radionuclide transfer in temperate environments, Technical Reports Series No. 364. International Atomic Energy Agency, Vienna
12. International Atomic Energy Agency (IAEA) (2010) Transfer to animals. In: Handbook of parameter values for the prediction of radionuclide transfer in terrestrial and freshwater environments. Technical Reports Series No. 472. International Atomic Energy Agency, Vienna
13. International Atomic Energy Agency (IAEA) (2009) Quantification of radionuclide transfer in terrestrial and freshwater environments for radiological assessments, IAEA-TECDOC-1616. International Atomic Energy Agency, Vienna
14. Beresford NA, Howard BJ, Voigt G (2007) Transfer of radionuclides to food producing animals. J Environ Radioactiv 98:1–3
15. Calmon P, Fesenko S, Voigt G, Linsley G (2009) Quantification of radionuclide transfer in terrestrial and freshwater environments. J Environ Radioactiv 100:671–674
16. Fesenko S, Isamov N, Howard BJ, Voigt G, Beresford NA, Sanzharova N (2007) Review of Russian language studies on radionuclide behaviour in agricultural animals: Part 1 gut absorption. J Environ Radioactiv 98:85–103
17. Fesenko S, Howard BJ, Isamov N, Voigt G, Beresford NA, Sanzharova N, Barnett CL (2007) Review of Russian language studies on radionuclide behaviour in agricultural animals: part 2 transfer to milk. J Environ Radioactiv 98:104–136
18. Fesenko S, Isamov N, Howard BJ, Beresford NA, Barnett CL, Sanzharova N, Voigt G (2009) Review of Russian language studies on radionuclide behaviour in agricultural animals: part 3 transfer to muscle. J Environ Radioactiv 100:215–231
19. Fesenko S, Howard BJ, Isamov N, Beresford NA, Barnet CL, Sanzharova N, Voigt G (2009) Review of Russian language studies on radionuclide behaviour in agricultural animals: part 4 poultry. J Environ Radioactiv 100:815–822
20. Velasco H, Juri Ayub J (2009) Root uptake: tropical and subtropical environments. In: International Atomic Energy Agency (ed) IAEA quantification of radionuclide transfer in terrestrial and freshwater environments for radiological assessments, IAEA-TECDOC-1616. International Atomic Energy Agency, Vienna, pp 207–238
21. Beresford NA, Howard BJ (1991) The importance of soil adhered to vegetation as a source of radionuclides ingested by grazing animals. Sci Total Environ 107:237–254
22. Fesenko SV, Alexakhin RM, Balonov MI, Bogdevich IM, Howard BJ, Kashparov VA, Sanzharova NI, Panov AV, Voigt G, Zhuchenka Yu (2007) An extended critical review of twenty years of countermeasures used in agriculture after the Chernobyl accident. Sci Total Environ 383:1–24
23. Vidal M, Rigol A, Gil-Garcia CJ (2009) Soil-radionuclide interactions. In: International Atomic Energy Agency (ed) Quantification of radionuclide transfer in terrestrial and freshwater environments for radiological assessments, IAEA-TECDOC-1616. International Atomic Energy Agency, Vienna, pp 71–102
24. Gil-Garcia C, Rigol A, Vidal M (2009) New best estimates for radionuclide solid-liquid distribution coefficients in soils, Part 1: radiostrontium and radiocaesium. J Environ Radioactiv 100:690–696
25. Vandenhove H, Gil-Garcia C, Rigol A, Vidal M (2009) New best estimates for radionuclide solid-liquid distribution coefficients in soils, Part 2: naturally occurring radionuclides. J Environ Radioactiv 100:697–703
26. Gil-Garcia C, Tagami K, Uchida S, Rigol A, Vidal M (2009) New best estimates for radionuclide solid-liquid distribution coefficients in soils, Part 3: miscellany of radionuclides (Cd, Co, Ni,

Zn, I, Se, Sb, Pu, Am, and others). J Environ Radioactiv 100: 704–715

27. Ward GM, Johnson JE, Stewart HF (1965) Cesium-137 passage from precipitation to milk. In: Klement AW (ed) Proceedings of the second conference on radioactive fallout from nuclear weapons tests. National Technical Information Service, Springfield, pp 703–710
28. Ward GM, Johnson JE (1965) The caesium-137 content of beef from dairy and feed lot cattle. Health Phys 11:95–100
29. Brown J, Simmonds JR (1995) Farmland: a dynamic model for the transfer of radionuclides through terrestrial foodchains, NRPB-R275. National Radiological Protection Board, Chilton
30. Müller H, Pröhl G (1993) ECOSYS-87 – a dynamic-model for assessing radiological consequences of nuclear accidents. Health Phys 64:232–252
31. Nuclear Regulatory Commission (NRC) (1977) Calculation of annual doses to man from routine releases of reactor effluents for the purpose of evaluating compliance with 10 CFR 50. Appendix I. Regulatory Guide 1.109. Nuclear Regulatory Commission, Office of Standards Development, Washington, DC
32. Yu C, Zielen AJ, Cheng J-J, LePoire DJ, Gnanapragasam E, Kamboj S, Arnish J, Wallo A III, Williams WA, Peterson H (2001) Users manual for RESRAD version 6. ANL/EAD-4. US Department of Energy, Office of scientific and technical information, Oak Ridge, http://web.ead.anl.gov/resrad/documents/resrad6.pdf
33. Ward GM, Johnson JE (1989) Assessment of milk transfer coefficients for use in prediction models of radioactivity transport. Sci Total Environ 85:287–294
34. Voigt G (1988) The transfer of ^{60}Co from feed into vitamin B12 in cow liver, milk and beef. J Environ Radioactiv 8:209–215
35. Beresford NA, Crout NMJ, Mayes RW, Howard BJ, Lamb CS (1998) Dynamic distribution of radioisotopes of cerium, ruthenium and silver in sheep tissues. J Environ Radioactiv 38: 317–338
36. Beresford NA, Howard BJ, Mayes RW, Lamb CS (2007) The transfer of radionuclides from saltmarsh vegetation to sheep tissues and milk. J Environ Radioactiv 98:36–49
37. Ward GM, Johnson JE (1986) Validity of the term transfer coefficient. Health Phys 50:411–414
38. Solheim Hansen H, Hove K (1991) Radiocaesium bioavailability: transfer of Chernobyl and tracer radiocaesium to goat milk. Health Phys 60:665–673
39. Howard BJ, Mayes RW, Beresford NA, Lamb CS (1989) Transfer of radiocaesium from different environmental sources to ewes and suckling lambs. Health Phys 57:579–586
40. Beresford NA, Lamb CS, Mayes RW, Howard BJ, Colgrove PM (1989) The effect of treating pastures with bentonite on the transfer of Cs-137 from grazed herbage to sheep. J Environ Radioactiv 9:251–264
41. Assimakopoulos PA, Ioannides KG, Karamanis DT, Pakou AA, Stamoulis KC, Mantizios AS, Nikolaou E (1993) Radiocaesium transfer to sheep's milk as a result of soil ingestion. Sci Total Environ 136:13–24
42. Belli M, Blas M, Capra E, Drigo A, Menegon S, Piasentier E, Sansone U (1993) Ingested soil as a source of ^{137}Cs to ruminants. Sci Total Environ 136:243–249
43. Belli M, Sansone U, Piasentier E, Capra E, Drigo A, Menegon S (1993) ^{137}Cs transfer coefficients from fodder to cow milk. J Environ Radioactiv 21:1–8
44. Voigt G, Müller H, Paretzke HG, Bauer T, Rohtmoser G (1993) ^{137}Cs transfer. Health Phys 65:141–146
45. Beresford NA, Mayes RW, Cooke AI, Barnett CL, Howard BJ, Lamb CS, Naylor GPL (2000) The importance of source dependent bioavailability in determining the transfer of ingested radionuclides to ruminant derived food products. Environ Sci Technol 34:4455–4462
46. Assimakopoulos PA, Ioannides KG, Karamanis DT, Pakou AA, Stamoulis KC, Mantizios AS, Nikolaou E (1994) Variation of the transfer coefficient for radiocaesium transport to sheep's milk during a complete lactation period. J Environ Radioactiv 22:63–75
47. Beresford NA, Mayes RW, Barnett CL, Howard BJ (2007) The transfer of radiocaesium to ewes through a breeding cycle – an illustration of the pitfalls of the transfer coefficient. J Environ Radioactiv 98:24–35
48. Beresford NA, Mayes RW, Barnett CL, MacEachern PJ, Crout NMJ (1998) Variation in the metabolism of radiocaesium between individual sheep. Radiat Environ Biophys 37: 277–281
49. Beresford NA, Mayes RW, Barnett CL, Lamb CS (2002) Dry matter intake- a generic approach to predict the transfer of radiocaesium to ruminants? Radioprotection colloques 37:373–378
50. Comar CL, Wasserman RH, Lengemann FW (1966) Effect of dietary calcium on secretion of strontium into milk. Health Phys 12:1–6
51. Comar CL, Wasserman RH, Twardock AR (1961) Secretion of calcium and strontium into milk. Health Phys 7:69–80
52. Comar CL (1966) Radioactive materials in animals – entry and metabolism. In: Russell RS (ed) Radioactivity and human diet. Pergamon, Oxford, pp 127–156
53. Lengemann FW (1963) Overall aspects of calcium and strontium absorption. In: Wasserman RH (ed) Transfer of calcium and strontium across biological membranes. Academic, London, pp 85–96
54. Agricultural Research Council (ARC) (1980) The nutrient requirements of ruminant livestock, Technical review by an Agricultural Research Council Working Group. CAB International, Wallingford
55. Howard BJ, Beresford NA, Mayes RW, Hansen HS, Crout NMJ, Hove K (1997) The use of dietary calcium intake of dairy ruminants to predict the transfer coefficient of radiostrontium to milk. Radiat Environ Biophys 36:39–43
56. Van den Hoek J (1989) European research on the transfer of radionuclides to animals – an historical perspective. Sci Total Environ 85:17–27
57. Beresford NA, Mayes RW, Hansen HS, Crout NMJ, Hove K, Howard BJ (1998) Generic relationship between calcium

intake and radiostrontium transfer to milk of dairy ruminants. Radiat Environ Biophys 3:129–131

58. Kazakov VS, Demidchik EP, Astakhova LN (1992) Thyroid cancer after Chernobyl. Nature 359:21
59. Likhtarev IA, Gulko GM, Sobolev BG, Kairo IA, Chepurnoy NI, Prohl G, Henrichs K (1994) Thyroid dose assessment for the Chernigov region (Ukraine): estimation based on ^{131}I thyroid measurements and extrapolation of the results to districts without monitoring. Radiat Environ Biophys 33:149–166
60. Jackson D, Jones SR (1991) Reappraisal of environmental countermeasures to protect members of the public following the Windscale Nuclear Reactor accident 1957. In: CEC (ed) Proceedings of a seminar on comparative assessment of the environmental impact of radionuclides released during three major nuclear accidents: Kyshtym, Windscale, Chernobyl, vol II. Commission of the European Communities, EUR 13574, Luxembourg, pp 1015–1040
61. Howard BJ, Beresford NA, Barnett CL, Fesenko S (2009) Gastrointestinal fractional absorption of radionuclides in adult domestic ruminants. J Environ Radioactiv 100:1069–1078
62. Crout NMJ, Beresford NA, Mayes RW, MacEachern PJ, Barnett CL, Lamb CS, Howard BJ (2000) A model of radioiodine transfer to goat milk incorporating the influence of stable iodine. Radiat Environ Biophys 39:59–65
63. Voigt G, Kiefer P (2007) Stable and radioiodine concentrations in cow milk: dependence on iodine intake. J Environ Radioactiv 98:218–227
64. Assimakopoulos PA, Ioannides KG, Pakou AA (1989) A study of radiocaesium contamination and decontamination of sheep's milk. Sci Total Environ 85:279–285
65. Howard BJ, Beresford NA, Mayes RW, Lamb CS (1993) Transfer of ^{131}I to sheep milk from vegetation contaminated by Chernobyl fallout. J Environ Radioactiv 19:155–161
66. Howard BJ, Voigt G, Beresford NA (2001) Transfer to food producing animals. Parts A and B. In: Van der Stricht E, Kirchmann R (eds) Radioecology, radioactivity and ecosystems. Fortemps, Liege, pp 58–169
67. Howard BJ, Beresford NA, Gashchak S, Arkhipov A, Mayes RW, Caborn J, Strømann G, Wacker L (2007) The transfer of $^{239/240}$Pu to cow milk. J Environ Radioactiv 98:191–204
68. Beresford NA, Mayes RW, Crout NMJ, MacEachern PJ, Dodd BA, Barnett CL, Lamb CS (1999) The transfer of cadmium and mercury to sheep tissues. Environ Sci Technol 33: 2395–2402
69. Ennis ME, Johnson JE, Ward GM, Voigt GM (1988) A specific activity effect in the metabolism of technetium. Health Phys 54:157–160
70. Ennis ME, Ward GM, Johnson JE, Boamah KN (1988) Transfer coefficients of selected radionuclides to animal products, 2. Hen eggs and meat. Health Phys 54:167–170
71. Ennis ME, Ward GM, Johnson JE, Boamah KN (1989) Technetium metabolism by lactating goats. Health Phys 57:321–330
72. Johnson JE, Ward GM, Ennis ME Jr, Boamah KN (1988) Transfer coefficients of selected radionuclides to animal products, 1. Comparison of milk and meat from dairy cows and goats. Health Phys 54:161–166
73. Jones BEV (1989) Technetium metabolism in goats and swine. Health Phys 57:331–336
74. Bondietti EA, Garten CTJ (1988) Transfer of I-131 and Tc-95 m from pasture to goat milk. In: Desmet G, Myttenaere C (eds) Technetium in the environment. Elsevier Applied Science, London, pp 339–347
75. Von Wiechen A, Heine K, Hagemeister H (1983) A contribution to the question of the transfer of technetium-99 into milk. Atomkernenerg Kerntech 42:199–200, In German
76. Voigt G, Henrichs K, Prohl G, Paretzke HG (1988) Measurements of transfer coefficients for ^{137}Cs, ^{60}Co, ^{22}Na, ^{131}I, and ^{95m}Tc from feed into milk and beef. Radiat Environ Biophys 27:143–152
77. Howard BJ, Beresford NA, Barnett CL, Fesenko S (2009) Radionuclide transfer to animal products: revised recommended transfer coefficient values. J Environ Radioactiv 100:263–273
78. Galeriu D, Crout NMJ, Melintescu A, Beresford NA, Peterson SR, van Hees M (2001) A metabolic derivation of tritium transfer coefficients in animal products. Radiat Environ Biophys 40:325–334
79. Galeriu D, Melintescu A, Beresford NA, Takeda H, Crout NMJ (2009) The dynamic transfer of ^{3}H and ^{14}C in mammals: a proposed generic model. Radiat Environ Biophys 48:29–45
80. Smith J, Beresford NA, Shaw GG, Moberg L (2005) Radioactivity in terrestrial ecosystems. In: Smith JT, Beresford NA (eds) Chernobyl, catastrophe and consequences. Springer/Praxis, Chichester, pp 81–137
81. Howard BJ, Beresford NA, Barnett CL, Fesenko S (2009) Quantifying the transfer of radionuclides to food products from domestic farm animals. J Environ Radioactiv 100: 767–773
82. Fesenko SV, Colgan PA, Sanzharova NI, Lissianski KB, Vazquez C, Guardans R (1997) The dynamics of the transfer of caesium-137 to animal fodder in areas of Russia affected by the Chernobyl accident and resulting from the consumption of milk and milk products. Radiat Prot Dosim 69: 289–298
83. Wright SM, Smith JT, Beresford NA, Scott WA (2003) Monte-Carlo prediction of changes in areas of west Cumbria requiring restrictions on sheep following the Chernobyl accident. Radiat Environ Biophys 42:41–47
84. International Atomic Energy Agency (IAEA) (2006) Fundamental safety principles. IAEA Safety Standards Series No. SF-1. International Atomic Energy Agency, Vienna, p 37
85. International Commission on Radiological Protection (ICRP) (2007) Recommendations of the international commission on radiological protection, vol 2–3, ICRP publication 103; Annals of the ICRP 37. Pergamon, Oxford
86. International Union of Radioecology (IUR) (2000) Dose and effects in non-human systems. Summary of the work of the action group of IUR. IUR Report 01. International Union of Radioecology Osteras, Norway. Available at www.iur-uir.org

87. Nuclear Energy Agency (NEA) (2007) Scientific issues and emerging challenges for radiological protection: report of the expert group on the implications of radiological protection sciences, No. 6167. Nuclear Energy Agency, France. ISBN 978-92-64-99032-6
88. Howard BJ, Beresford NA, Andersson P, Brown JE, Copplestone D, Beaugelin-Seiller K, Garnier-Laplace J, Howe PD, Oughton D, Whitehouse P (2010) Protection of the environment from ionising radiation in a regulatory context—an overview of the PROTECT coordinated action project. J Radiol Prot 30:195–214
89. United States Department of Energy (USDOE) (2002) A graded approach for evaluating radiation doses to aquatic and terrestrial biota, Technical standard DOE-STD-1153-2002, Modules 1-3. United States Department of Energy, Washington, DC
90. Copplestone D, Bielby S, Jones SR, Patton D, Daniel CP, Gize I (2001) Impact assessment of ionising radiation on wildlife, R&D Publication 128. Environment Agency and English Nature, Bristol
91. Copplestone D, Wood MD, Bielby S, Jones SR, Vives i Batlle J, Beresford NA (2003) Habitat regulations for stage 3 assessments: radioactive substances authorisations. R&D Technical Report P3- 101/SP1a. Environment Agency, Bristol
92. Keum D-K, Jun I, Lim K-M, Choi Y-H (in press) External dose conversion coefficients to assess the radiological impact of an environmental radiation on aquatic and terrestrial animals. J Nucl Sci Tech
93. Beaugelin-Seiller K (2006) EDEN version 2 – User's manual. Report ECRE/06-29. IRSN/DEI, pp 38
94. Nedveckaite T, Filistovic V, Marciulioniene D, Remeikis V, Beresford NA (2007) Exposure of biota in the cooling pond of Ignalina NPP: hydrophytes. J Environ Radioactiv 97:137–147
95. Garisto NC, Cooper F, Fernandes SL (2008) No-effect concentrations for screening assessment of radiological impacts on non-human biota. NWMO TR-2008-02. Nuclear Waste Management Organisation, Toronto. Available at www.nwmo.ca
96. Chouhan SL, Yankovich TL, Davis PA (2009) Environmental radionuclide concentrations below which non-human biota experience no effects. Radioprotection 44:107–114
97. Environment Canada, Health Canada (2003) Releases of radionuclides from nuclear facilities (Impact on non-human biota): final report May 2003, Priority substances list assessment report: Canadian Environmental Protection Act 1999. Minister of Public Works and Government Services Canada, Ottawa
98. Williams C (ed) (2004) Framework for assessment of environmental impact (FASSET) of ionising radiation in European ecosystems. J Radiol Prot 24
99. Howard BJ, Larsson C-M (2008) The ERICA integrated approach and its contribution to protection of the environment from ionising radiation. J Environ Radioactiv 99:1361–1363
100. International Commission on Radiation Protection (ICRP) (2008) Environmental protection: the concept and use of reference animals and plants, ICRP Publication 108; Ann. ICRP 38. International Commission on Radiation Protection, Oxford
101. International Atomic Energy Agency (IAEA) (in-press) Modelling radiation exposure and radionuclide transfer for nonhuman species. Report of the biota working group of the EMRAS theme 3: environmental modelling for radiation safety (EMRAS) programme. International Atomic Energy Agency, Vienna. http://www-ns.iaea.org/downloads/rw/projects/emras/final-reports/biota-final.pdf
102. Brown JE, Alfonso B, Avila R, Beresford NA, Copplestone D, Pröhl G, Ulanovsky A (2008) The ERICA tool. J Environ Radioactiv 99:1371–1383
103. International Atomic Energy Agency (IAEA) (2001) Generic models for use in assessing the impact of discharges of radioactive substances to the environment, IAEA Safety Reports Series 19, STI/PUB/1102. International Atomic Energy Agency, Vienna
104. Vives i Batlle J, Wilson RC, Watts SJ, McDonald P, Craze A (2009) Derivation of allometric relationships for radionuclides in marine phyla. Radioprotection 44:47–52
105. Avila R, Beresford NA, Agüero A, Broed R, Brown J, Iospje M, Robles B, Suañez A (2004) Study of the uncertainty in estimation of the exposure of non-human biota to ionizing radiation. J Radiol Prot 24:A105–A122
106. Olyslaegers G (in-press) Modelling radiation exposure and radionuclide transfer for non-human species, Appendix IV DosDiMEco. Report of the biota working group of the EMRAS theme 3: environmental modelling for radiation safety (EMRAS) programme. International Atomic Energy Agency, Vienna. http://www-ns.iaea.org/downloads/rw/projects/emras/final-reports/biota-final.pdf
107. Beresford NA, Wright SM, Sazykina T (2001) Reference Arctic organisms. Contract deliverable under EPIC Inco-copernicus Project ICA2-CT02000-10032. Centre for Ecology and Hydrology: Merlewood, Grange-over-Sands. https://wiki.ceh.ac.uk/display/rpemain/EPIC+reports
108. Strand P, Beresford NA, Avila R, Jones SR, Larsson C-M (eds) (2001) Overview of radiation exposure pathways relevant for the identification of candidate reference organisms. FASSET Deliverable 1, Identification of candidate reference organisms from a radiation exposure pathways perspective. EC 5th framework programme Contract No. FIGE-CT-2000-00102. Including Appendix 1 and 2 on ecological characteristics of European terrestrial (App.1) and aquatic (App. 2) ecosystems. Norwegian Radiation Protection Authority, Østerås. https://wiki.ceh.ac.uk/display/rpemain/FASSET+reports
109. Beresford NA, Barnett CL, Howard BJ, Scott WA, Brown JE, Copplestone D (2008) Derivation of transfer parameters for use within the ERICA tool and the default concentration ratios for terrestrial biota. J Environ Radioactiv 99:1393–1407
110. United States Department of Energy (USDOE) (2004) RESRAD-BIOTA: a tool for implementing a graded approach to biota evaluation: User Guide Version 1, DOE/EH-0676. United States Department of Energy, Springfield

111. Hosseini A, Thørring H, Brown JE, Saxén R, Ilus E (2008) Transfer of radionuclides in aquatic ecosystems – default concentration ratios for aquatic biota in the ERICA tool. J Environ Radioactiv 99:1408–1429
112. Brown JE (ed) (2003) Environmental protection from ionising contaminants in the Arctic: Project ICA2-CT-2000-10032. Second annual report for EPIC (01.11.2001 – 31.10.2002). Norwegian radiation Protection Authority, Østerås. https://wiki.ceh.ac.uk/display/rpemain/EPIC+reports
113. Brown J, Strand P, Hosseini A, Borretzen P (eds) (2003) Handbook for assessment of the exposure of biota to ionising radiation from radionuclides in the environment. FASSET Deliverable 5. Norwegian Radiation Protection Authority, Østerås. https://wiki.ceh.ac.uk/display/rpemain/FASSET+reports
114. Vives i Batlle J, Balonov M, Beaugelin-Seiller K, Beresford NA, Brown J, Cheng J-J, Copplestone D, Doi M, Filistovic V, Golikov V, Horyna J, Hosseini A, Howard BJ, Jones SR, Kamboj S, Kryshev A, Nedveckaite T, Olyslaegers G, Pröhl G, Sazykina T, Ulanovsky A, Vives Lynch S, Yankovich T, Yu C (2007) Inter-comparison of unweighted absorbed dose rates for non-human biota. Radiat Environ Biophys 46:349–373
115. Vives i Batle J, Beaugelin-Seiller K, Beresford NA, Copplestone D, Horyna J, Hossein, A, Johansen M, Kamboj S, Keum D-K, Kurosawa N, Newsome L, Olyslaegers G, Vandenhove H, Ryufuku S, Vives Lynch S, Wood MD, Yu C (2011) The estimation of absorbed dose rates for non-human biota: an extended inter-comparison. Radiat Environ Biophys 50:231–251
116. Beresford NA, Barnett CL, Brown J, Cheng J-J, Copplestone D, Filistovic V, Hosseini A, Howard BJ, Jones SR, Kamboj S, Kryshev A, Nedveckaite T, Olyslaegers G, Saxen R, Sazykina T, Vives i Batlle J, Vives-Lynch S, Yankovich T, Yu C (2008) Intercomparison of models to estimate radionuclide activity concentrations in non-human biota. Radiat Environ Biophys 47:491–514
117. Beresford NA, Barnett CL, Brown JE, Cheng J-J, Copplestone D, Gashchak S, Hosseini A, Howard BJ, Kamboj S, Nedveckaite T, Olyslaegers G, Smith JT, Vivesi Batlle J, Vives-Lynch S, Yu C (2010) Predicting the radiation exposure of terrestrial wildlife in the Chernobyl exclusion zone: an international comparison of approaches. J Radiol Prot 30:341–373
118. Beresford NA, Hosseini A, Brown JE, Cailes C, Beaugelin-Seiller K, Barnett CL, Copplestone D (2010) Assessment of risk to wildlife from ionising radiation: can initial screening tiers be used with a high level of confidence? J Radiol Prot 30:265–284
119. Yankovich TL, Vives i Batlle J, Vives-Lynch S, Beresford NA, Barnett CL, Beaugelin-Seiller K, Brown JE, Cheng J-J, Copplestone D, Heling R, Hosseini A, Howard BJ, Kamboj S, Kryshev AI, Nedveckaite T, Smith JT, Wood MD (2010) An international model validation exercise on radionuclide transfer and doses to freshwater biota. J Radiol Prot 30:299–340
120. Higley KA, Domotor SL, Antonio EJ (2003) A kinetic-allometric approach to predicting tissue radionuclide concentrations for biota. J Environ Radioactiv 66:61–74
121. International Atomic Energy Agency (IAEA) (in-preparation) Handbook of transfer parameters for the prediction of radionuclide transfer to wildlife. Technical Report Series No. XX. International Atomic Energy Agency, Vienna
122. Garnier-Laplace J, Della-Vedova C, Andersson P, Copplestone D, Cailes C, Beresford NA, Howard BJ, Howe P, Whitehouse P (2010) A multi-criteria weight of evidence approach for deriving ecological benchmarks for radioactive substances. J Radiol Prot 30:215–233
123. Yankovich TL, Beresford NA, Wood MD, Aono T, Andersson P, Barnett CL, Bennett P, Brown J, Fesenko S, Hosseini A, Howard BJ, Johansen M, Phaneuf M, Tagami K, Takata H, Twining J, Uchida S (2010) Whole-body to tissue concentration ratios for use in biota dose assessments for animals. Radiat Environ Biophys 49:549–565
124. Peters RH (1983) The ecological implications of body size. Cambridge University Press, Cambridge
125. Brown JH, West GB, Enquist BJ (2000) Patterns and processes, causes and consequences scaling in biology. In: Brown JH, West GB (eds) Scaling in biology. Oxford University Press, New York, pp 1–24
126. West GB, Brown JH, Enquist BJ (2000) The origin of universal scaling laws in biology. In: Brown JH West GB, West JH (eds) Scaling in biology. Oxford University Press, New York, pp 7–112
127. Macdonald CR (1996) Ingestion rates and radionuclide transfer in birds and mammals on the Canadian shield, Report TR-722 COG-95-551. AECL, Ontario
128. Cherry RD, Heyraud M (1991) Polonium-210 and lead-210 in marine organisms: allometric relationships and their significance. In: Kershaw PJ, Woodhead DS (eds) Radionuclides in the study of marine processes. Elsevier Applied Science, London, pp 309–318
129. Vives i Batlle J, Wilson RC, McDonald P (2007) Allometric methodology for the calculation of biokinetic parameters for marine biota. Sci Total Environ 388:256–269
130. Sheppard SC (2001) Toxicants in the environment: bringing radioecology and ecotoxicology together. In: Brechignac F, Howard BJ (eds) Radioactive pollutants impact on the environment. EDP Sciences, France, pp 63–74
131. Brown J, Børretzen P, Dowdall M, Sazykina T, Kryshev I (2004) The derivation of transfer parameters in the assessment of radiological impacts to arctic marine biota. Arctic 57:279–289
132. Higley KA, Bytwerk DP (2007) Generic approaches to transfer. J Environ Radioactiv 98:4–23
133. Beresford NA, Broadley MR, Howard BJ, Barnett CL, White PJ (2004) Estimating radionuclide transfer to wild species – data requirements and availability for terrestrial ecosystems. J Radiol Prot 24:A89–A103

134. Enquist BJ, Kerkhoff AJ, Stark SC, Swenson NG, McCarthy MC, Price CA (2007) A general integrative model for scaling plant growth, carbon flux, and functional trait spectra. Nature 449:218–222
135. Niklas KJ (2006) Plant allometry, leaf nitrogen and phosphorus stoichiometry and interspecific trends in annual growth rates. Ann Bot 97:155–163
136. Higley KA (2010) Estimating transfer parameters in the absence of data. Radiat Environ Biophys 49:645–656
137. Nagy KA (2001) Food requirements of wild animals: predictive equations for free-living mammals, reptiles and birds. Nutr Abstr Rev B 71:1R–12R
138. Willey NJ (2010) Phylogeny can be used to make useful predictions of soil-to-plant transfer factors for radionuclides. Radiat Environ Biophys 49:613–623
139. Broadley MR, Willey NJ, Wilkins J, Baker AJM, Mead A, White P (2001) Phylogenetic variation in heavy metal accumulation in angiosperms. New Phytol 152:9–27
140. Willey NJ, Tang S, McEwen A, Hicks S (2010) The effects of plant traits and phylogeny on soil-plant transfer of 99Tc. J Environ Radioactiv 101:757–766
141. Jeffree RA, Oberhansli F, Teyssie J-L (2010) Phylogenetic consistencies among chondrichthyan and teleost fishes in their bioaccumulation of multiple trance elements from seawater. Sci Total Environ 408:3200–3210
142. Tagami K, Uchida S (2010) Can elemental composition of crop leaves be used to estimate radionuclide transfer to tree leaves? Radiat Environ Biophys 49(4): 583–590
143. Railsback LB (2003) An earth scientist's periodic table of the elements and their ions. Geology 31:737–740
144. Tyler G (2004) Ionic charge, radius, and potential control root/soil concentration ratios of fifty cationic elements in the organic horizon of a beech (*Fagus sylvatica*) forest podzol. Sci Total Environ 329:231–239
145. Galeriu D, Beresford NA, Melintescu A, Avila R, Crout NMJ (2003) Predicting tritium and radiocarbon in wild animals. In: International Atomic Energy Agency (ed) International conference on the protection of the environment from the effects of ionizing radiation (IAEA-CN-109), Stockholm, 6–10 October 2003. International Atomic Energy Agency, Vienna, pp 186–189
146. Smith KL, Robinson CA, Jones SR, Vives i Batlle J, Norris S (2008) Assessment of impact to non-human biota from a generic waste repository in the UK. In: Strand P, Brown J, Jølle T (eds) Proceeding of international conference on radioecology and environmental radioactivity: oral and poster presentations, part 2, Bergen, 15–20 June 2008. Norwegian Radiation Protection Authority, Østerås, pp 415–418
147. Kryshev AI (2006) ^{90}Sr in fish: a review of data and possible model approach. Sci Total Environ 370:182–189
148. Allott R, Copplestone D (2008) Update on habitats assessments for England and Wales. National Dose Assessment Working Group. Paper 13-04. http://www.ndawg.org/documents/Paper13-04.pdf
149. Smith JT, Kudelsky AV, Ryabov IN, Hadderingh RH (2000) Radiocaesium concentration factors of Chernobyl-contaminated fish: a study of the influence of potassium and "blind" testing of a previously developed model. J Environ Radioactiv 48:359–369
150. Smith JT, Sasina NV, Kryshev AI, Belova NV, Kudelsky AV (2009) A review and test of predictive models for the bioaccumulation of radiostrontium in fish. J Environ Radioactiv 100:950–954
151. Copplestone D, Brown JE, Beresford NA (2010) Considerations for the integration of human and wildlife radiological assessments. J Radiol Prot 30:283–297
152. Gillett AG, Crout NMJ, Absalom SM, Wright SM, Young SD, Howard BJ, Barnett CL, McGrath SP, Beresford NA, Voigt G (2001) Temporal and spatial prediction of radiocaesium transfer to food products. Radiat Environ Biophys 40:227–235
153. Tipping E (1994) WHAM – a chemical equilibrium model and computer code for waters, sediments and soils incorporating a discrete-site electrostatic model of ion-binding by humic substances. Comput Geosci 20:973–1023
154. Takata H, Aono T, Tagami K, Uchida S (2010) Concentration ratios of stable elements for selected biota in Japanese estuarine areas. Radiat Environ Biophys 49:591–601
155. Uchida S, Tagami K, Shang ZR, Choi YH (2009) Uptake of radionuclides and stable elements from paddy soil to rice: a review. J Environ Radioactiv 100:739–745

Books and Reviews

Eisenbud M, Gesell T (1997) Environmental radioactivity from natural, industrial and military sources, 4th edn. Academic, San Diego

International Atomic Energy agency (IAEA) (2005) Protection of the environment from the effects of ionizing radiation. In: International Atomic Energy agency (IAEA) (ed) Protection of the environment from the effects of ionizing radiation: Proceedings of an International conference, Stockholm, 6–10 October 2003. International Atomic Energy agency, Vienna

Kathren RL (1984) Radioactivity in the environment: sources, distribution and surveillance. Harwood Academic Publishers, Switzerland

Scott EM (ed) (2003) Modelling radioactivity in the environment. Elsevier, Amsterdam

Smith JT, Beresford NA (eds) (2005) Chernobyl, catastrophe and consequences. Springer/Praxis, Chichester

van der Stricht E, Kirchmann R (eds) (2001) Radioeology, radioactivity & ecosytems. International Union of Radioecology (IUR), FORTEMPS, Liege, Belgium

Whicker FW, Schulz V (1982) Radioecology: nuclear energy and the environment, vol 1&2. CRC Press, Boca Raton

Radiation Effects on Caribou and Reindeer

BLISS L. TRACY
Radiation Protection Bureau, Health Canada 6604C,
Ottawa, ON, Canada

Article Outline

Glossary
Definition of the Subject
Introduction
The Species *Rangifer terandus*
The Sources of Radionuclides in the Environment of Caribou/Reindeer
Radionuclide Levels in Lichens
Radionuclide Levels in *Rangifer tarandus*
Effects in *Rangifer tarandus*
Impacts on Human Populations
Future Directions
Bibliography

Glossary

Absorbed dose The amount of radiation energy absorbed per unit mass of tissue.

Alpha radiation A form of ionizing radiation, consisting of tightly bound particles containing two protons and two neutrons. This is the least penetrating form of ionizing radiation and can be stopped by a sheet of paper or by a layer of skin.

Becquerel (Bq) The SI (Système International) unit of activity, defined as one disintegration per second.

Beta radiation A form of ionizing radiation consisting of positively or negatively charged electrons. This is a moderately penetrating form of ionizing radiation and can be stopped by 1 mm of aluminum.

Equivalent dose The absorbed dose multiplied by a radiation weighting factor, equal to one for beta and gamma radiation, 20 for alpha radiation, etc.

Gamma radiation A form of ionizing radiation consisting of very-high-frequency electromagnetic waves. This is the most penetrating form of ionizing radiation and requires 10 cm of lead to significantly reduce its intensity.

Gray (Gy) The SI (Système International) unit of absorbed dose, equal to 1 J/kg. One Gray = 1,000 milligrays (mGy).

Half-life The time required for the number of radioactive atoms in a substance to decrease to one half of its initial value.

Ionizing radiation Radiation capable of stripping electrons from atoms as it passes through matter, hence creating ions. Normally emitted by the nuclei of radioactive elements.

Nuclear fission A process in which a heavy nucleus (e.g., ^{235}U) splits into two lighter fragments with the release of a large amount of energy. The resulting fragments are usually radioactive.

Radioactivity The property of the nuclei of certain elements to undergo disintegration, with the emission of one or more particles of ionizing radiation.

Radionuclide A radioactive substance characterized by the number of protons and neutrons in its nuclei. For example, the radionuclide cesium 137 (^{137}Cs) is characterized by 55 protons and 82 neutrons, giving a total mass of 137.

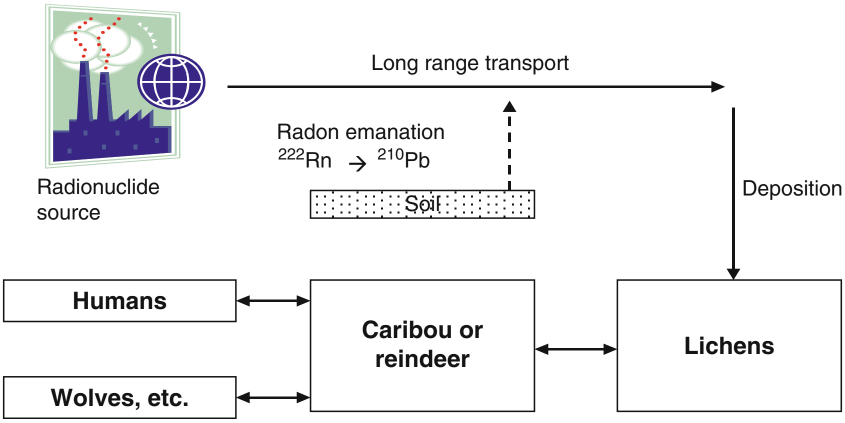
Sievert (Sv) The SI (Système International) unit of equivalent dose, equal to absorbed dose in grays multiplied by a radiation weighting factor. One Sievert = 1,000 millisieverts (mSv).

Definition of the Subject

Since the 1960s, extensive studies have been conducted on the uptake and retention of radionuclides in the lichen → caribou → human food chain. The reported concentrations of radioactivity in caribou and reindeer have been higher than in any other large terrestrial mammal. The results from these studies need to be brought together to address a fundamental question: Are these elevated levels of radioactivity having any measurable impact on the health of the animals? A supplementary question to be answered is: What is the impact on human populations?

Introduction

It has been known since the early 1960s that caribou and reindeer are particularly vulnerable to radioactive fallout and other airborne contaminants. To understand why this is so, one must examine the position and role of these animals in the ecosystem.



Radiation Effects on Caribou and Reindeer. Figure 1
A schematic diagram of the radioecology of caribou and reindeer

Figure 1 provides a simplified view of the caribou/reindeer ecosystem, emphasizing those features responsible for the uptake of radionuclides.

Sources of radioactivity in the environment may be natural (e.g., radon emanation from the ground) or artificial (e.g., a nuclear reactor or nuclear explosion). Radioactivity released into the atmosphere as a fine aerosols may be transported hundreds or thousands of kilometers from its original source by prevailing winds and global circulation patterns.

Gravitational settling and washout by precipitation eventually cause the radioactive aerosols to be deposited on the ground and on vegetation surfaces.

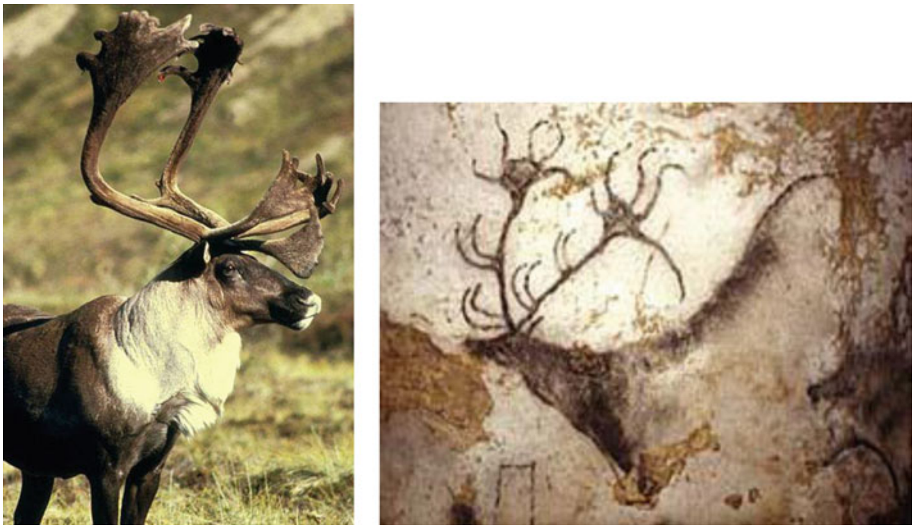
One vegetation class of particular importance are the lichens, which have large surface areas and no root systems. They depend on the atmosphere for many of their nutrients and thus are efficient collectors of a wide range of airborne pollutants, including radionuclides. Furthermore, lichens are not renewed annually, like grasses or deciduous leaves, but live for decades, which means they can accumulate contaminants over long periods of time.

Caribou and reindeer (*Rangifer tarandus*) are the only mammalian species that can subsist on a diet of lichens during the long winter months. This accounts for their reproductive success and proliferation in harsh northern climates and also for their vulnerability to ecosystem contamination. In order to obtain sufficient nourishment, they must graze the lichens over a large area and thus become integrators of radioactive fallout, both in space and in time. Caribou and reindeer serve as food for both humans and natural predators, which are often in competition with one another.

The relationship between *Homo sapiens* and *Rangifer tarandus* is a very ancient one, as attested to by cave paintings and artifacts found in southern Europe from 17,000 years ago (Fig. 2). Arctic and subarctic peoples, including the Sami of Fennoscandia and the Nenets of northern Siberia, have herded reindeer since the Middle Ages. Reindeer are raised for their meat, hides, antlers, and, to a lesser extent, for milk and for use as draft animals or beasts of burden.

The ancestors of the North American Indians migrated across the Bering Strait land bridge 10,000–12,000 years ago in search of game animals such as caribou. Today, caribou are still the most important dietary source of protein in many communities of Alaska and northern Canada. In the late nineteenth century, reindeer were introduced into Alaska as semi-domesticated livestock as a means of providing a livelihood for native peoples. Descendants of that herd populate several game farms in Canada today.

Recent studies are beginning to show that populations of caribou are on the decline. The reasons for this decline are still poorly understood. It is likely due to a combination of factors – climate change, natural predators, overhunting by humans, disruptions of migration routes by highways and oil pipelines from northern development, and, last but not least, environmental contaminants. It is debatable whether



Radiation Effects on Caribou and Reindeer. Figure 2
A modern reindeer and a cave painting from 17,000 years ago

exposures to environmental contaminants alone can explain the decline, but contaminants may interact with other environmental stressors and weaken the overall ability of the animals to survive.

To gain further understanding of one major class of contaminants, this entry traces the origins and transport of both natural and artificial radionuclides in the environment. It documents their uptake and concentration levels in lichens and caribou/reindeer. The question of possible effects on the animals is addressed in light of the latest knowledge on radiation effects. The impacts on human populations who depend on caribou/reindeer are discussed as well.

The Species *Rangifer terandus*

Both caribou and reindeer are members of the same species, *Rangifer tarandus*, which in turn belongs to the *cervidae* or deer family. There are a number of distinct subspecies of *Rangifer*, but caribou and reindeer do not fall neatly into these categories. In fact, the variability within these two types is greater than the difference between them. The main distinction is probably etymological. Essentially, "caribou" is a New World term for the *Rangifer tarandus* that are native to North America and "reindeer" is an Old World term for the ones that inhabit northern Europe today. The term "caribou" comes, through French, from Mi'kmaq *qalipu*, meaning "snow shoveler," and refers to the animals' habit of pawing through the snow for food [1]. The English word "reindeer" comes from the Norwegian *reinsdyr* [old Norse *hreinn* + *dyr* ("animal")].

Since 1961, *Rangifer* has been divided into two major groups: tundra with six subspecies and woodland with three subspecies.

Tundra *Rangifer tarandus*

- Arctic Reindeer (*R. tarandus eogroenlandicus*), an extinct subspecies found until 1900 in eastern Greenland
- Peary Caribou (*R. tarandus pearyi*), found in the northern islands of the Nunavut and the Northwest Territories of Canada
- Svalbard Reindeer (*R. tarandus platyrhynchus*), found on the Svalbard islands of Norway, the smallest subspecies
- Mountain/Wild Reindeer (*R. tarandus tarandus*), found in the Arctic tundra of Eurasia, including the Fennoscandia peninsula of northern Europe
- Porcupine Caribou or Grant's Caribou (*R. tarandus granti*), found in Alaska, the Yukon, and the Northwest Territories of Canada. Very similar to *R. tarandus groenlandicus* described below, and

probably better regarded as a junior synonym of that subspecies

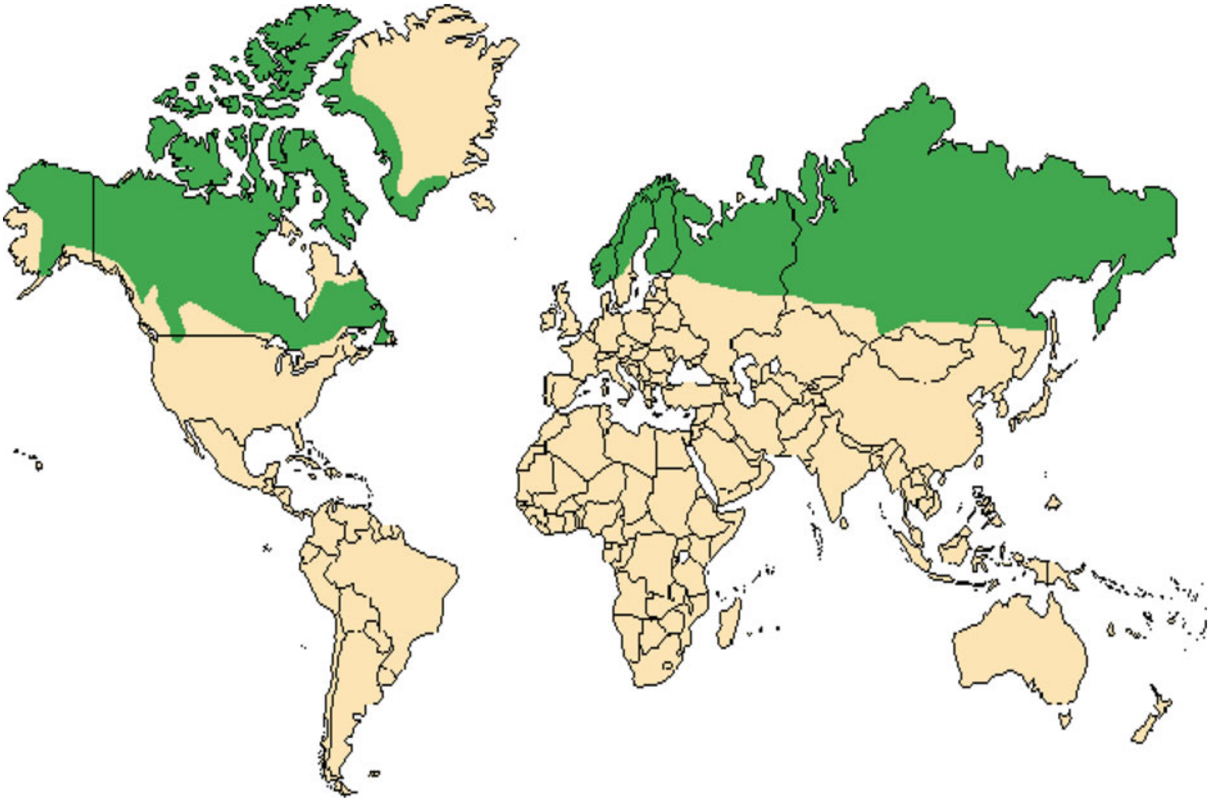
- Barren-ground Caribou (*R. tarandus groenlandicus*), found in Nunavut and the Northwest Territories of Canada and in western Greenland

Woodland *Rangifer tarandus*

- Finnish Forest Reindeer (*R. tarandus fennicus*), found in the wild in only two areas of Fennoscandia – Finnish/Russian Karelia and a small population in central south Finland
- Migratory Woodland Caribou (*R. tarandus caribou*), or Forest Caribou, once found in the North American taiga (boreal forest) from Alaska to Newfoundland and Labrador and as far south as New England, Idaho, and Washington. Woodland Caribou have disappeared from most of their original southern range and are considered threatened where they remain, with the notable exception of the Migratory Woodland Caribou of northern Quebec and Labrador
- Queen Charlotte Islands Caribou (*R. tarandus dawsoni*), which became extinct at the beginning of the twentieth century

Rangifer tarandus is widely distributed throughout the circumpolar region and range as far south as 46° North Latitude (Fig. 3). It was once believed that *Rangifer* evolved in Europe and Asia, and moved into North America during the time of the land bridge across the Bering Strait (10,000–20,000 years ago). However, *Rangifer* bones dating back 1.5 million years have been found in the Fort Selkirk area of the Yukon. It now seems likely that *Rangifer tarandus* evolved in North America, moved to Europe and Asia, and then moved back into North America across the land bridge.

It is an oversimplification to state that all caribou are wild and all reindeer domesticated. Reindeer are not considered fully domesticated, as they generally roam



Radiation Effects on Caribou and Reindeer. Figure 3
World distribution of *Rangifer tarandus* shown in green

free on pasture grounds and they have never been bred in captivity. In traditional nomadic herding, reindeer herders migrate with their herds between coast and inland areas according to an annual migration route, and herds are keenly tended. Furthermore, truly wild reindeer do exist in Norway. They number about 25,000 animals and are found in 23 more or less separated areas in the mountainous southern part of the country. A total of 4,817 animals, or about 20% of the population, was harvested in 2005. Also, a diminutive subspecies of wild reindeer occurs throughout the unglaciated parts of the Svalbard Archipelago. Although no exact count has been carried out, these reindeer are believed to number several thousands.

Rangifer are ruminants, having a four-chambered stomach. They eat mainly lichens in winter, especially reindeer moss. However, they also eat the leaves of willows and birches, as well as sedges and grasses. Some populations of the North American caribou migrate the furthest of any terrestrial mammal, traveling up to 5,000 km in 1 year. They normally travel about 20–60 km per day while migrating, but can run at speeds of 60–80 km/h. During the spring migration, smaller herds will group together to form larger herds of 50,000–500,000 animals but during autumn migrations, the groups become smaller and the caribou begin to mate.

Rangifer tarandus is the only species of deer in which both sexes have antlers. Males are significantly larger than females, with weights ranging from 100 to 200 kg and occasionally exceeding 300 kg. Female weights vary from 80 to 120 kg. Male caribou compete for access to females during the fall rut, which occurs in October and early November. During this time, males may engage in battles that leave them injured and exhausted. Dominant males restrict access to small groups of 5–15 females. Males stop feeding during this time and lose much of their body reserves. Calves are born the following May or June. After 45 days, the calves are able to graze and forage but continue suckling until the following autumn and become independent from their mothers.

Although populations of *Rangifer* have been relatively stable in the past, there is some evidence of declining caribou herds, especially in Canada's Northwest Territories. A Territorial government survey released in September 2009 indicated that the Bathurst herd population had decreased from 128,000 to 32,000 animals between 2003 and 2009 [2]. In the mid-1980s, this herd numbered nearly 500,000 animals. The Cape Bathurst and Bluenose herds showed similar decreases over this period. The causes of these declines are not well understood. A number of possible causes have been put forward, including climate change, natural predators, overhunting by humans, disruptions of migration routes by highways and oil pipelines from northern development, and environmental contaminants.

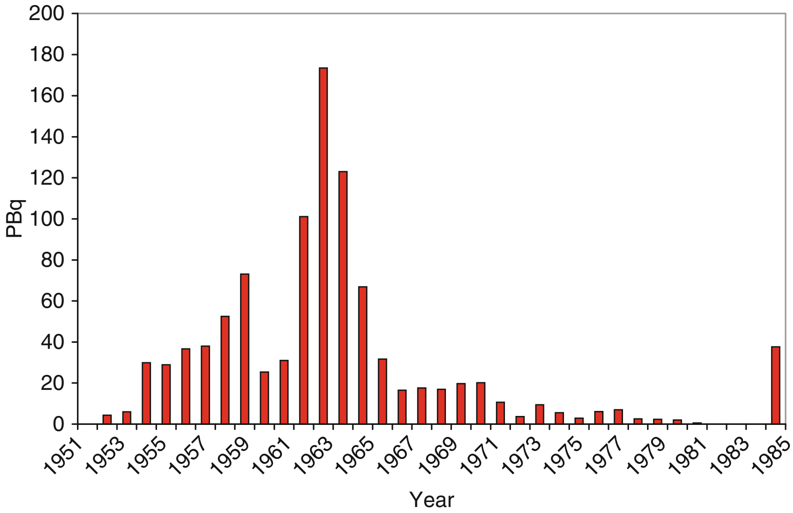
The Sources of Radionuclides in the Environment of Caribou/Reindeer

Artificial Radionuclides

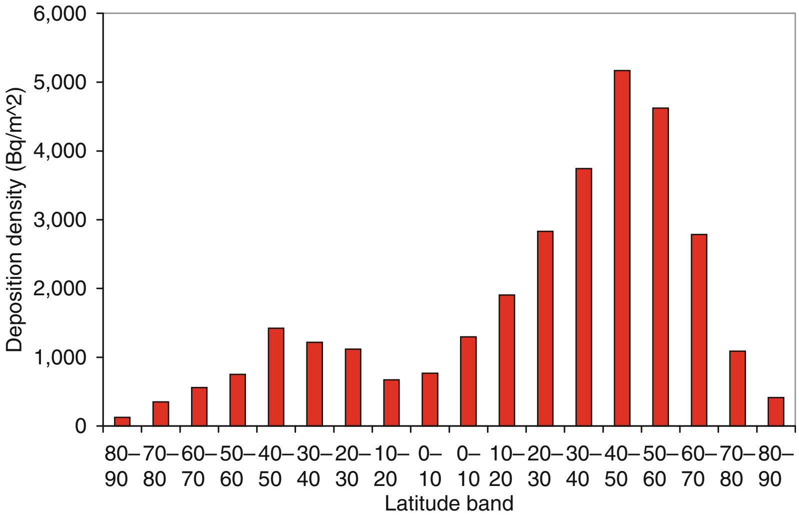
Two of the most important artificial radionuclides are cesium 137 (^{137}Cs) and strontium 90 (^{90}Sr). Both are produced with high yields in nuclear fission; both have half-lives of about 30 years, and thus persist in the environment for long periods of time; and both are highly radiotoxic. Cesium is a chemical analogue of potassium and is readily taken up by all living organisms where it becomes more or less uniformly distributed throughout the body of the organism. Strontium is an analogue of calcium and becomes concentrated in the bones and teeth of animals where it may remain for years.

The major source of these radionuclides in the global environment has been fallout from the atmospheric testing of nuclear weapons. Figure 4 shows the annual global deposition of ^{137}Cs since 1951 [3]. Fallout declined sharply after 1963, when the United States, the Soviet Union, and the United Kingdom signed a treaty banning the testing of nuclear weapons in the atmosphere. France and China, non-signatories of the treaty, continued some atmospheric testing up until 1980, leading to a continuing input of fission products at a reduced level. A large amount of ^{137}Cs and ^{134}Cs were injected into the atmosphere following the accident in 1986 at the Chernobyl nuclear power station in the Ukraine. Fallout radiocesium was significantly augmented over large areas of Europe and Asia during the year of the accident. The less-volatile ^{90}Sr was not emitted to the same degree.

Figure 5 shows the cumulative distribution of ^{137}Cs deposition worldwide, integrated over 10° latitude



Radiation Effects on Caribou and Reindeer. Figure 4
Annual global deposition of ^{137}Cs from 1951 to 1986. The large spike in 1986 is due to the Chernobyl accident. The units are Petabecquerels = 10^{15} Bq (Based on data from UNSCEAR [3])



Radiation Effects on Caribou and Reindeer. Figure 5
Cumulative deposition density of ^{137}Cs from 1951 to 1985, averaged over each 10° latitude band. The contribution from the Chernobyl accident in 1986 is not included, since this deposition was very nonuniform (Based on data from UNSCEAR [3])

bands [3]. Fallout was higher in the Northern Hemisphere than the Southern Hemisphere, and higher in temperate zones than polar zones. However, fallout was still significant from Latitude 50° to 70° N, where most of the world's caribou and reindeer reside.

Natural Radionuclides

The natural radioactive elements uranium and thorium are present at average levels of 20–40 Bq/kg in the earth's crust, although higher levels are observed in zones of significant mineralization. These elements normally remain locked in rocks and soils, until they are brought to the surface by mining operations. However, an important exception occurs in nature. One of the decay products of both uranium and thorium is radon gas, which readily diffuses out of the soil. Normally, it disperses quickly in outdoor air but can build up to hazardous levels in buildings and underground mines. One of the decay products of radon 222 (^{222}Rn) is lead 210 (^{210}Pb), which is important in the caribou/reindeer ecosystem. Lead 210 in turn decays to polonium 210 (^{210}Po), a highly toxic, alpha-emitting radionuclide. The reader may recall that ^{210}Po was the poison used in the famous Russian spy case of 2006 [4]. The partial decay scheme leading to these radionuclides is shown below. Half-lives are shown in brackets.

^{238}U (4.5 billion years) → intermediate products → ^{222}Rn (3.824 days) → short-lived products (∼minutes) → ^{210}Pb (21 years) → ^{210}Bi (5.01 days) → ^{210}Po (138.4 days) → ^{206}Pb (stable)

With a half-life of 22 years, ^{210}Pb can be transported over long distances and then be deposited on lichens, where it builds up over decades. The average residence time of ^{210}Pb in air is about 10 days [5], which is sufficient for global transport. Once deposited, ^{210}Po soon grows into equilibrium with ^{210}Pb and the two radionuclides are ingested by caribou and reindeer which graze on the lichens.

Radionuclide Levels in Lichens

Lichens are composite organisms consisting of a symbiotic association of a fungus (the mycobiont) with a photosynthetic partner (the photobiont or phycobiont), usually either a green alga or cyanobacterium. Lichens do not have roots and do not need to tap continuous reservoirs of water like most higher plants, thus they can grow in locations impossible for most plants, such as bare rock, sterile soil, or sand. Lichens occur in some of the most extreme environments on earth – arctic tundra, hot deserts, rocky coasts, and toxic slag heaps. Lacking a root system, their primary source of most nutrients is the air. Therefore, elemental levels in lichens often reflect the accumulated composition of ambient air. Lichens are widespread and can live for decades; however, they are vulnerable to environmental pollution and do not fare well in most urban environments.

Several factors account for the high concentrations of radionuclides found in Arctic lichens.

- They have a large surface area to mass ratio. Total deposited radioactivity depends on surface area, so this leads to a high radionuclide concentration per unit mass.
- Their lack of a root system makes them efficient at absorbing trace elements from air.
- With a lifetime of decades, they can accumulate pollutants over a long period of time.
- In northern environments, the growing season is short and vegetation turnover is generally slow. This leads to long-term retention times of pollutants in lichens.

^{137}Cs

As early as 1959, Gorham [6] noted the ability of lichens to concentrate radionuclides in the environment. Hofmann et al. [7] stated that this is due to the absorption of cesium by the mycobiont, that is, the fungal part of the lichen, in order to satisfy its potassium requirements.

Hanson [8] reported ^{137}Cs levels in lichens from four locations in Alaska between 1962 and 1979. He noted a dramatic increase during the peak of nuclear testing 1962–1963, reaching levels of 25–30 nCi kg^{-1} (900–1,100 Bg kg^{-1}) dry weight. He determined a caribou intake of 4.5–5.0 kg dry lichens day^{-1}, which would lead to a radiocesium intake of up to 5,000 Bq day^{-1}. The ^{137}Cs content of lichens declined only very slowly over subsequent years. In a separate experiment, he determined the retention half-time of cesium on lichens to be 5 ± 2 years, quite close to currently accepted values.

Enormous increases in radiocesium concentrations in lichens were observed in Europe after the Chernobyl accident of 1986. Hofmann et al. [7] measured the enhancement factors of 23–547 for ^{137}Cs in Austrian lichens before and after Chernobyl, showing the extreme vulnerability of this plant. White et al. [9] observed a more modest increase of 75% for lichens in the Fairbanks area of Alaska, which was less affected by Chernobyl fallout.

In 1980, Hutchison-Benson, Svoboda, and Taylor [10, 11] measured ^{137}Cs concentrations in vegetation and top soil along a south–north transect in Canada from 58° to 82° N. They found concentrations in lichens up to 2,000 Bg kg^{-1} dry weight – five times higher than in vascular plants from the same locations. The concentrations in all plants followed a bell-shaped distribution curve with latitude, reaching a maximum between 60° and 70° N, farther north than the maximum in ^{137}Cs fallout shown in Fig. 5. The authors attributed this northward shift to the slower vegetation turnover rates at higher latitudes. They found that the effective ^{137}Cs half-life in the lichen species *Alectoria nigricans* increased from about 4 years at 50° N to 8 years at 82° N.

The above findings have important implications for the study of radiocesium in *Rangifer*. First of all, the latitude band 60°–70° N is where the majority of world's caribou and reindeer herds are found – in Alaska, Canada, Fennoscandia, and Siberia. Second, the effective half-life of ^{137}Cs in the caribou/reindeer ecosystem is driven by its retention time in vegetation, particularly lichens.

^{210}Pb and ^{210}Po

The detection of enhanced ^{137}Cs uptake by lichens led some investigators to suspect that naturally occurring airborne radionuclides, such as ^{210}Pb and ^{210}Po, might also be concentrated by this mechanism. In the 1960s, Holtzman [12] measured these two radionuclides in a suite of biota and human specimens from Alaska. He found ^{210}Pb levels in Alaskan lichens of 420–2,500 Bq kg^{-1} dry weight, whereas Beasley and Palmer [13] reported 170 Bq kg^{-1} wet weight for a composite lichen sample from Alaska.

Kauranen and Miettinen [14] measured ^{210}Po and ^{210}Pb concentrations in lichens collected from various locations in Finland between 1961 and 1967. Their results are reproduced here in Table 1 to illustrate the

Radiation Effects on Caribou and Reindeer. Table 1 ^{210}Po and ^{210}Pb content (Bq kg^{-1} dry weight) of lichen samples (*Cladonia alpestris*) from different parts of Finland from [14]. (Units in the original publication were pCi g^{-1} dry weight)

Date of collection	Locality	^{210}Po	^{210}Pb	^{210}Po/^{210}Pb
	Lapland			
August 1961	Inari	–	310	–
July 1964	Inari	250	270	0.92
July 1966	Enontekiö	350	380	0.93
July 1966	Inari	210	210	0.98
	Average	270	290	0.92
	Southern Finland			
October 1964	Tuusula	220	270	0.80
July 1965	Pielisjärvi	220	240	0.92
August 1965	Porvoo	180	200	0.91
October 1966	Helsinki	180	230	0.82
October 1966	Virolahti	210	260	0.84
October 1966	Loppi	170	190	0.90
April 1967	Helsinki	200	260	0.78
April 1967	Virolahti	260	300	0.86
May 1967	Loppi	200	230	0.85
	Average	200	240	0.85

consistency of these concentrations over time and space. It is notable that the ^{210}Po/^{210}Pb ratios are only slightly less than 1.00, indicating that the ^{210}Po in lichens is nearly in secular equilibrium with ^{210}Pb.

Jaworowski [15] measured ^{210}Pb in lichens along a south–north transect from 41° to 77° N and found higher values in the Arctic (200–500 Bq kg^{-1}dry weight) than in the temperate zone (30–120 Bq kg^{-1}). He also observed an increase in ^{210}Pb levels in samples from Polish glaciers dated between 1952 and 1965. A likely explanation for this increase is anthropogenic input from nuclear weapons testing, specifically from the reaction ^{208}Pb($2n$, γ)^{210}Pb in the intense neutron flux created by the nuclear explosion.

Persson [16] reported ^{210}Pb concentrations in archived lichen samples from Sweden, collected between 1882 and 1972. Although there were considerable fluctuations in the data, no consistent trends were apparent during this long time period. More intensive sampling from 1962 to 1972 gave a range of ^{210}Pb concentrations between 100 and 300 Bq kg^{-1} dry weight.

Table 2 shows more recent results from Skuterud et al. [17] on three radionuclides in Norwegian lichens – ^{90}Sr, ^{137}Cs, and ^{210}Po. The ^{210}Po concentrations were not notably different from the values obtained in Finland 40 years earlier.

In short, the ^{210}Pb content of Arctic lichens varies from about 100 to 300 Bq kg^{-1} dry weight, with the ^{210}Po almost in secular equilibrium with ^{210}Pb. These levels are not insignificant, compared to the ^{137}Cs levels from fallout, especially when one considers that ^{210}Po, an alpha emitter, could be 10–20 times more radiotoxic than ^{137}Cs for the same level of activity. Apart from a small temporary increase during the period of intensive nuclear weapons testing in the 1960s, the ^{210}Pb levels appear to have been constant in time over the past 130 years.

Radionuclide Levels in *Rangifer tarandus*

^{137}Cs

Monitoring of radiocesium in the caribou and Inuit of Alaska began in 1962. At that time, Hanson and Palmer [18] observed a close correlation between body burdens in animals and in people over the next 3 years. In a later paper, Hanson [8] established that there was a clear annual cycle in the ^{137}Cs levels in caribou flesh, with low values occurring in the fall months and highest values in the spring month after feeding

Radiation Effects on Caribou and Reindeer. Table 2 Activity concentrations of ^{90}Sr, ^{137}Cs, and ^{210}Po in the living part of *Cladonia arbuscula* at different locations in Vågå and Østre Namdal (Bq kg^{-1} DM)

District	Location	Sampling date	^{90}Sr	^{137}Cs	^{210}Po
Vågå	1	August 2001	9.1 ± 1.1	3470 ± 280	121 ± 4
	2	August 2001	12.1 ± 1.5	1,460 ± 120	-
	3	August 2001	-	1,115 ± 96	104 ± 3
	7	August 2001	22.6 ± 2.8	1,910 ± 160	192 ± 6
	11	August 2001	12.0 ± 1.5	3,060 ± 250	70 ± 2
	12	September 2002	-	1,890 ± 160	212 ± 5
Østre namdal	D-2	September 2001	6.0 ± 0.7	1,930 ± 160	119 ± 3
	D-4	September 2001	9.7 ± 1.2	1,380 ± 120	171 ± 4
	0	July 2003	7.9 ± 1.0	3,180 ± 250	111 ± 3
	1	July 2003	-	3,510 ± 270	154 ± 3
	S-1	July 2003	5.5 ± 0.7	1,430 ± 110	-
	J-3	July 2003	-	1,570 ± 120	151 ± 4

on lichens over the winter. The highest value in caribou meat occurred during the period 1964–1972, with spring maxima reaching about 7,000 Bq kg^{-1} dry weight during some years.

During the early 1960s, Norway began investigations of possible high exposures to ^{137}Cs in the Sami population of Finnmark, due to substantial fallout registered in this region and due to the habits of the reindeer breeders [19]. They reported that radiocesium levels in reindeer meat had decreased from a high of 3,000 Bq kg^{-1} wet weight in 1966 to a value of 350 Bq kg^{-1} by 1983. This implies a T_{eff} of about 7 years for ^{137}Cs in reindeer meat. Rissanen and Rahola [20] have published similar data for reindeer in Finland.

The Chernobyl accident of 1986 introduced a huge flux of radiocesium fallout into northwestern Europe, which had a profound effect on the lichen → reindeer → food chain. Fallout in some parts of Scandinavia exceeded 60,000 Bq m^2. Åhman and Åhman [21] reported levels of 20,000 Bq kg^{-1} wet weight in Swedish reindeer in the year of the accident, and values as high as 80,000 Bq kg^{-1} the following year. The results showed a strong seasonality, with the lowest values in September and highest from January through April. Åhman and Åhman followed the decline in ^{137}Cs concentrations from 1986 to 1992, and derived effective halftimes of 2.3–5.0 years, depending on seasonality and locations. These halftimes are significantly lower than the values obtained from global nuclear weapons fallout.

The impact of Chernobyl fallout was much less in North America. Using the signature from ^{134}Cs (found in reactor emissions but not nuclear weapons fallout), Macdonald et al. [22] estimated that Chernobyl fallout contributed only an additional 10–20% to the radiocesium burdens in Canadian caribou herds. White et al. [12] predicted that the impact on Alaskan herds could have been as high as 75%.

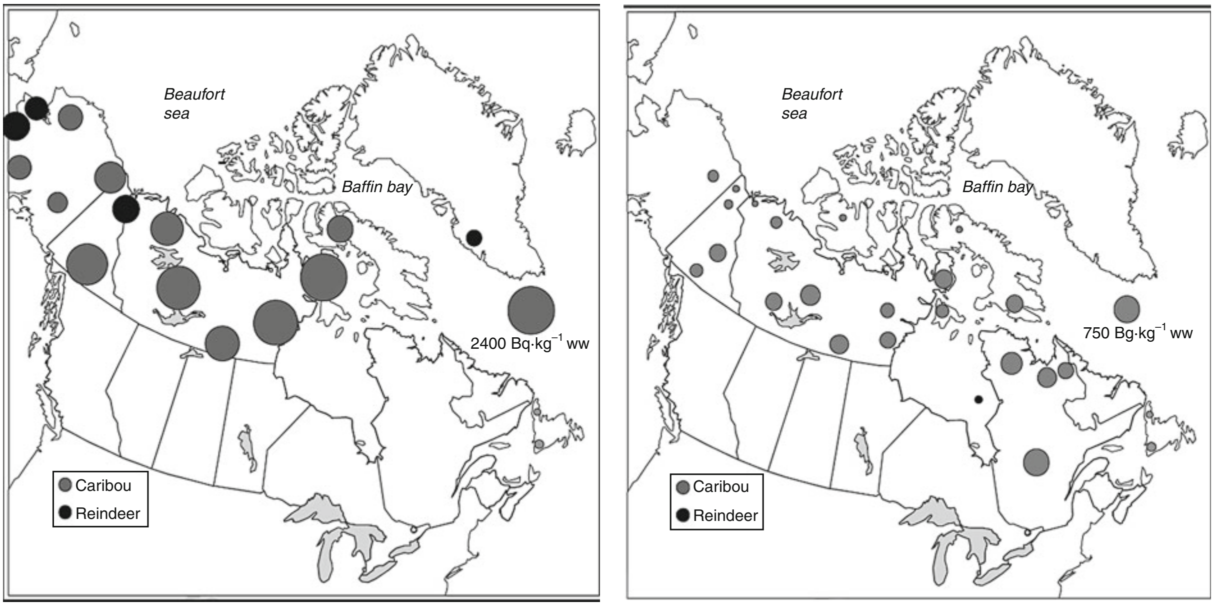
Macdonald et al. [22] recently published a comprehensive review of all radiocesium data in caribou and reindeer across Canada, Alaska, and Greenland over the 40-year period from 1960 to 2000. Canada began sampling of caribou meat in 1963 onward, with measurements in northern peoples following shortly thereafter. The program was largely terminated after 1969, but was restarted after the Chernobyl accident of 1986 and continued sporadically throughout the 1990s. Concentrations of ^{137}Cs in some Canadian caribou herds reached levels of 2,000–3,000 Bq kg^{-1} in the 1960s, with one value of 6,400 Bq kg^{-1}. At the time of the Chernobyl accident, there were still some Canadian caribou showing ^{137}Cs concentrations above 1,000 Bq kg^{-1}. By the late 1990s, levels had decreased to a few hundred Bq kg^{-1} at most.

Figure 6 compares radiocesium levels in the various herds and demonstrates how they changed over a 20-year period from the mid-1960s to the mid-1980s. Figure 7 shows the decrease in radiocesium levels from the height of nuclear testing to the late 1990s. A fit to the data yielded an effective half-life of 6.1 years – not out of line with other estimates. Rissanen and Rahola [23] derived ecological half-lives for ^{137}Cs in reindeer meat of 4.9 ± 1.1 years in western Russia, 8.3 ± 4.7 years in eastern Russia, 5.4 ± 1.1 years in Finland, and 7.2 ± 1.3 years in Norway.

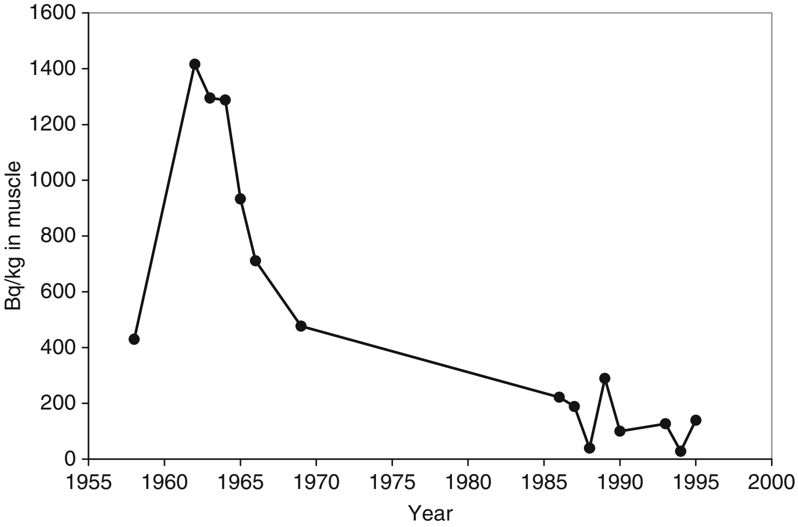
A similar review of radiocesium data in reindeer in the Eastern Hemisphere is available in the 1998 assessment report of the Arctic Monitoring and Assessment Programme (AMAP) [24]. Figure 8 shows the time trends for ^{137}Cs in caribou from 1960 to 1995 for five regions in Eurasia – Arctic Norway, Arctic Finland, Greenland, western Russia, and eastern Russia. The trends are quite similar to those in North America, with high values during the period of intensive atmospheric weapons testing in the 1960s followed by a gradual decline toward the present. The spike from the Chernobyl accident in 1986 is much more evident in the European data than the North American data. The map in Fig. 9 shows the average concentrations of ^{137}Cs in caribou and reindeer meat after 1990 at locations throughout the circumpolar region.

A number of studies have been conducted on the uptake, tissue distribution, and retention time of ^{137}Cs in caribou and reindeer. Hollemann et al. [25] injected experimental reindeer with radiocesium and also fed reindeer with lichens containing known radiocesium concentrations. They obtained a radiocesium absorption factor of 0.24 ± 0.031 for reindeer and retention times varying from 7 days in summer to 18 in winter. Applying these results to radiocesium levels in caribou in late spring, they estimated a lichen consumption rate of 3 kg dry weight per day.

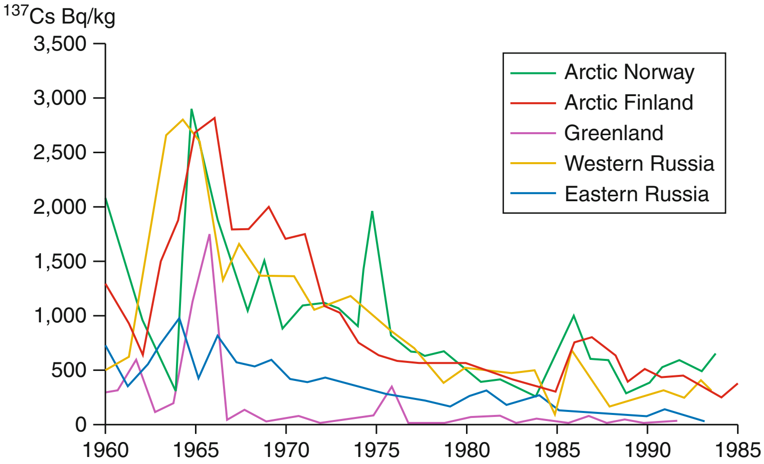
Macdonald et al. [22] reported results for ^{137}Cs concentrations in caribou liver, kidney, and heart.



Radiation Effects on Caribou and Reindeer. Figure 6
Spatial distribution of ^{137}Cs in North American caribou and reindeer during the 1960s (*left*) and the 1980s (*right*). The area of each circle is proportional to the pooled mean of all winter/spring concentrations for a given herd during the decade (From Macdonald et al. [22])



Radiation Effects on Caribou and Reindeer. Figure 7
Decline in ^{137}Cs concentrations in caribou muscle of Canadian herds from the 1960s to the 1990s. Each circle represents the geometric mean concentration of ^{137}Cs for all caribou herds in a given calendar year (Based on data from Macdonald et al. [22])



Radiation Effects on Caribou and Reindeer. Figure 8
Changes with time in activity concentration of ^{137}Cs in reindeer meat in Arctic Norway, Arctic Finland, Greenland, and Arctic Russia (From AMAP [24])

In general, ^{137}Cs concentrations were up to 30% higher in kidney than in muscle, whereas values in liver averaged about half those in muscle. In the three herds for which heart tissue was analyzed, the ^{137}Cs concentrations in heart were slightly lower than that in skeletal muscle. Macdonald et al. did not observe any significant correlation between ^{137}Cs concentration and age of the animals.

^{90}Sr

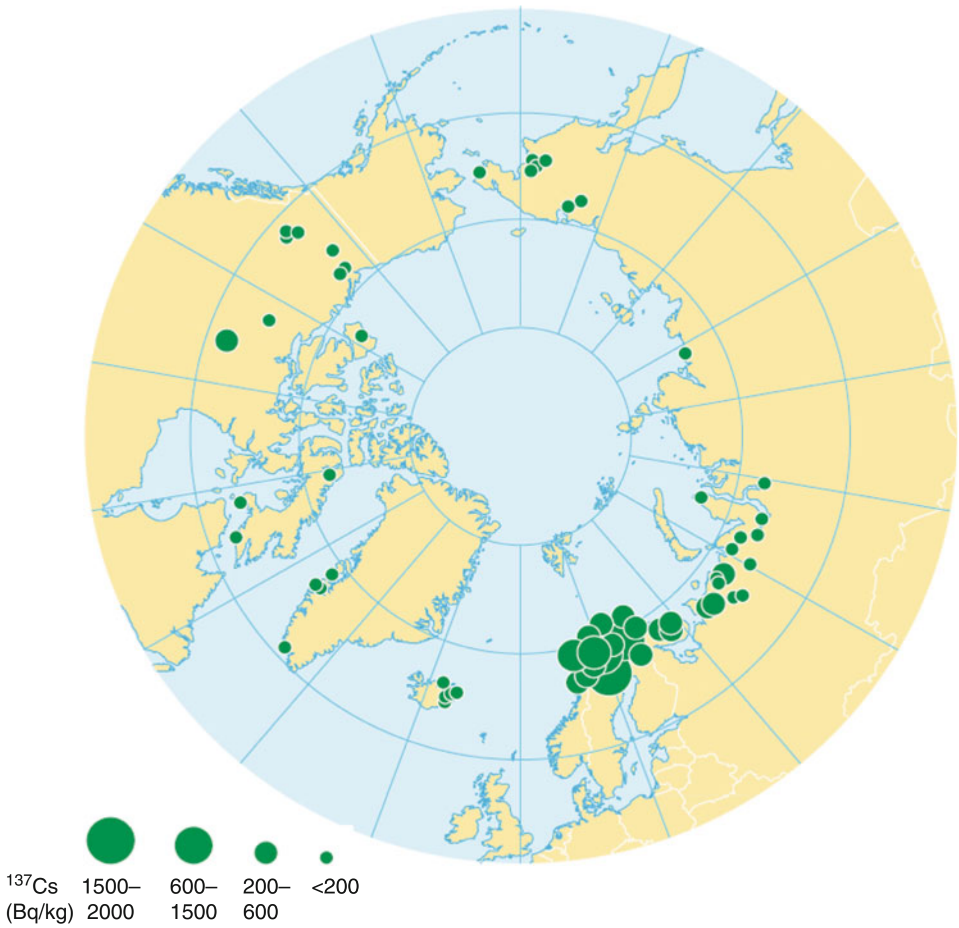
Much less has been reported on ^{90}Sr in caribou and reindeer. Certainly, ^{90}Sr was spread worldwide by global fallout in approximately the same amount as radiocesium. Like cesium, strontium can be transferred through the lichen → caribou → human food chain. However, strontium tends to be more mobile in the environment, and is not retained to the same degree in lichens. Being a chemical analog of calcium, strontium tends to accumulate in the bones of caribou and reindeer, which are not part of the human diet. Hence, less attention has been devoted to this radionuclide in Arctic food chains.

In 1962, Shulert [26] reported measurements of ^{90}Sr in the antlers and bones of Alaskan caribou. He obtained an average value of 7.7 Bq g^{-1} Ca for antlers, and values of 6.6 and 5.8 Bq g^{-1} Ca for the bones of two different caribou.

^{210}Pb and ^{210}Po

The accumulation of the naturally occurring radionuclides ^{210}Pb and ^{210}Po have been studied extensively in caribou and reindeer as well as in lichens. During the 1960s, ^{210}Pb and ^{210}Po were measured in Alaskan caribou by Beasely and Palmer [13], Blanchard and Kearney [27], and Blanchard and Moore [28]. They were also measured in Scandinavian reindeer by Kaurenen and Miettinen [14] and by Persson [16]. More recently, results have been reported for Canadian caribou by Macdonald et al. [29] and by Thomas et al. [30]. These data are summarized in Table 3.

It is apparent from Table 3 that there is a major redistribution of ^{210}Pb and ^{210}Po within the body of the animal, with ^{210}Pb concentrating primarily in bone and ^{210}Po in soft tissues, particularly liver and kidney. This is quite different from the situation in the lichens, where ^{210}Po is nearly in secular equilibrium with ^{210}Pb (Table 1). Data from Alaska and Scandinavia are quite similar to one another, but the Canadian values are consistently higher. This may be due to a number of factors. The air concentrations of ^{210}Pb can build up to higher levels over the large Canadian land mass than over Alaska or Scandinavia which are more influenced by coastal air currents. Also, the ^{137}Cs levels in caribou have



Radiation Effects on Caribou and Reindeer. Figure 9
Average activity concentrations of ^{137}Cs in reindeer meat after 1990 (From AMAP [24])

been consistently higher over Canada than Alaska (Figs. 4 and 5), which indicates differing deposition patterns and uptake by vegetation.

Major variations in ^{210}Pb and ^{210}Po concentrations were observed among different caribou herds in Canada. Macdonald et al. reported [29] ^{210}Pb concentrations in bone from the high Arctic Islands that were an order of magnitude less than in the central Arctic. A more detailed comparison is available from Thomas et al. [30], who studied complete tissue distributions of ^{210}Pb and ^{210}Po in 24 caribou taken by local hunters in each of the two Arctic communities – Baker Lake (Qamani'tuaq) in Nunavut Territory and Snowdrift (Łutsel K'e) in the Northwest Territories. Baker Lake lies in the tundra and relies on the Wager Bay and Kaminuriak herds; Snowdrift is located in boreal and scrub forest, and relies on caribou largely from the Beverly herd. Their results are summarized in Table 4. The concentrations in Baker Lake caribou were in most cases about twice those in Snowdrift caribou.

^{210}Pb and ^{226}Ra

Blanchard and Moore [28] reported measurements of ^{210}Pb and ^{226}Ra from five archived caribou bone samples collected in Canada before 1951. The ^{210}Pb concentrations in pre-1951 bone (240 ± 209 Bq kg^{-1}) were

Radiation Effects on Caribou and Reindeer. Table 3 A summary of ^{210}Pb and ^{210}Po measurements (Bq kg^{-1}) in caribou tissues. Standard deviations are given in brackets

Radionuclide and location	Muscle	Liver	Kidney	Bone
LEAD-210				
Alaska 1960s [13]	1.1	48 (27)	19 (3)	161 (35)
Alaska 1960s [28]	–	–	–	187 (85)
Alaska 1960s [27]	0.34 (0.12)	–	–	172 (67)
Finland 1960s [14]	0.22 (0.08)	25 (18)	–	169 (29)
Sweden 1970s [16]	0.48	26	10	192
Canada 1990s [29]	–	187 (168)	56 (21)	664 (308)
Canada 1990s [30]	–	136 (66)	71 (26)	793 (419)
POLONIUM-210				
Alaska 1960s [13]	13 (3)	171 (27)	145 (16)	140 (42)
Alaska 1960s [27]	7.5 (4.8)	–	–	131 (53)
Finland 1960s [14]	5.8 (3.9)	103 (59)	–	76 (2)
Canada 1990s [29]	–	431 (281)	227 (163)	387 (232)
Canada 1990s [30]	16 (9)	299 (147)	214 (108)	403 (266)

Radiation Effects on Caribou and Reindeer. Table 4 Concentrations of ^{210}Pb and ^{210}Po in the organs of caribou taken near two northern Canadian communities, from Thomas et al. [30]. Standard deviations are given in brackets

	Baker lake		Snowdrift	
Organ	Pb 210	Po 210	Pb 210	Po 210
Bone	1,023 (402)	531 (235)	562 (120)	274 (125)
Rumen	103 (24)	191 (39)	77 (16)	166 (54)
Kidney	84 (18)	259 (88)	57 (19)	168 (62)
Liver	158 (22)	374 (122)	113 (62)	224 (82)
Muscle		17 (8)		15 (5)
Placenta	2.7 (0.8)	26 (7)	1.9 (0.6)	11 (3)
Fetus	11 (7)	5.8 (2.0)	6.5 (1.9)	2.3 (0.6)
Fetal liver	25 (11)	16 (15)	10 (5)	35 (19)

not greatly different from the Alaskan data of 1965–66 (187 ± 85 Bq kg^{-1} in Table 3). These results confirm that the ^{210}Pb in the lichen → caribou food chain is likely of natural rather than anthropogenic origin. The concentrations of ^{226}Ra in the bone (6.5 ± 3.1 Bq kg^{-1}) are much lower, demonstrating that ^{210}Pb in bone is not supported by ^{226}Ra decay but originates from radon gas in the atmosphere.

^{40}K

Macdonald et al. [22] reported potassium 40 concentrations in Canadian caribou taken since 1986.

The mean concentrations and standard deviations in different organs were as follows:

Muscle	109 (22)
Liver	11 (26)
Kidney	122 (34)
Heart	93 (10)

The values were quite similar for all animals and organs, which is not surprising since potassium is homeostatically regulated in the bodies of animals.

Effects in *Rangifer tarandus*

The question now arises as to what are the health effects, if any, to the animals from these radionuclide levels. This question is addressed here in three steps. First, absorbed doses in organs and tissues are calculated based on the measured radionuclide concentrations. Then, the literature is surveyed to ascertain whether radiation effects might be expected at these levels of dose. Finally, a number of studies are examined from areas of high radioactive fallout, which suggest possible effects on caribou and reindeer.

Radiation Doses to Caribou and Reindeer

The primary factor governing radiation effects in living organisms is the absorbed dose, defined as the amount of radiation energy absorbed per kg of tissue. The SI unit for absorbed dose is the Gray (Gy), equal to 1 J of energy absorbed per kg. A quantity often used in human assessments is the equivalent dose which recognizes that some types of ionizing radiation are more detrimental than others. The equivalent dose in Sv is defined as absorbed dose in Gy multiplied by a radiation weighting factor equal to one for beta and gamma radiation, 20 for alpha radiation, 5 for protons, and 5–20 for neutrons.

A simple calculation of the absorbed dose in an organ or tissue relies can be carried out with the assumption that the radioactivity is uniformly distributed throughout an infinite medium. This assumption requires only that the range of the emitted radiation is small compared to the dimensions of the organ. It works well for alpha and beta radiation, but may slightly overestimate the dose from gamma radiation, some of which may escape from the tissue. According to this assumption, the amount of radiation energy absorbed per unit volume of tissue must equal the amount of energy generated per unit volume, or else there would be a net flow of energy from one part of the tissue to another. The energy generated per unit volume is readily calculated from the radionuclide concentration in Bq kg^{-1} and the expression for absorbed dose becomes

Dose (mGy year^{-1}) = radionuclide concentration (Bq kg^{-1}) × radiation energy (MeV) × 1.602×10^{13} (joules MeV^{-1}) × 86,400 (s day^{-1}) × 365.25 (d year^{-1}) × 1,000 (mGy Gy^{-1}).

The amount of radiation energy (MeV) per disintegration is given below for each of the relevant radionuclides considered here:

^{137}Cs	0.937 (mixed β and γ)
^{210}Pb	0.753 (mixed β and γ; includes decay of ^{210}Bi)
^{210}Po	5.305 (pure α)
^{40}K	0.570 (mixed β and γ)

The resulting doses to caribou organs and tissues are shown in Table 5. The measured distributions of ^{210}Pb and ^{210}Po in caribou tissues from Baker Lake (Table 4) were used, since these data have some of the highest values in Canadian caribou. Since ^{137}Cs concentrations in meat have varied widely over time and location, a range of values has been assumed for this radionuclide. The natural radionuclide ^{40}K was taken to be present in all tissues at about 100 Bq kg^{-1}. It was assumed here that all measured concentrations represent a steady state and are maintained at these levels for an entire year. Lead and polonium are grouped together in this table, since they are closely correlated. Cesium and potassium are grouped together because of their chemical similarity.

In the absence of fallout from nuclear weapons testing, the greatest contributors to doses in caribou have been the naturally occurring radionuclides – ^{210}Pb and ^{210}Po. The organs most greatly affected are bone, liver, and kidney, with doses to muscle an order of magnitude lower. It is notable that doses to placenta, fetus, and fetal liver are quite low, indicating that the lead and polonium activities are not passing on

Radiation Effects on Caribou and Reindeer. Table 5 Annual doses to caribou organs and tissues based on the measured concentrations of radionuclides

	Concentrations (Bq kg^{-1})		Doses (mGy year^{-1})		
Organ	Pb 210	Po 210	Pb 210	Po 210	Pb + Po
Bone	1,023	531	3.90	14.24	18.14
Kidney	84	259	0.32	6.95	7.27
Liver	158	374	0.60	10.03	10.63
Muscle	–	17	–	0.46	0.46
Placenta	2.7	26	0.01	0.70	0.71
Fetus	11	5.8	0.04	0.16	0.20
Fetal liver	25	16	0.10	0.43	0.52
	Cs 137	K 40	Cs 137	K 40	Cs + K
Muscle	80,000[a]	100	378.95	0.57	379.52
	20,000[b]	100	94.74	0.57	95.31
	7,000[c]	100	33.16	0.57	33.73
	3,000[d]	100	14.21	0.57	14.78
	1,000[e]	100	4.74	0.57	5.31
	200[f]	100	0.95	0.57	1.52

[a] Highest Swedish value from Chernobyl
[b] Typical Swedish value from Chernobyl
[c] Highest single value in North America
[d] Highest herd values in Canada, 1960s
[e] Highest herd values in Canada, 1980s
[f] Highest values in Canada, today (2010)

significantly to the developing fetus. The doses from ^{137}Cs range from a high of over 300 mGy year^{-1} (in Scandinavian reindeer after the Chernobyl accident) down to present-day values of about 1 mGy year^{-1} and approaching the doses from naturally occurring ^{40}K.

Dose–Effect Relationships in Animals

In recent years, there has been a paradigm shift in the protection of nonhuman biota from radiation. Before the early 1990s, it was generally assumed if radiation standards were set low enough to protect humans, then all other species would automatically be protected. The primary goal in radiation standard-setting is the protection of individual humans from stochastic effects such as cancer, where the risks are assumed to be directly proportional to radiation dose with no threshold (linear no threshold or LNT hypothesis). For nonhuman biota, it is the survival of the species (or the local subpopulation of that species) that is important. This places the focus on deterministic effects, such as cell death and reproductive failure, that will definitely occur but not below a certain dose threshold. This logic leads to the conclusion that other species can withstand more radiation than humans.

Today, there is an increased awareness of situations where the old paradigm may break down, either because there are no humans present in the immediate environment or because a certain species is supersensitive to the effects radiation at some stage in its life cycle. The lichen–caribou/reindeer ecosystem may represent one such situation. These animals wander over large stretches of remote territory which may be heavily contaminated by radioactive fallout.

Humans may not be immediately exposed to these high-contamination levels. In the previous section, it was seen how radionuclide concentrations can build up to high levels in caribou and reindeer.

In the past 2 decades, there has been a considerable effort to establish dose criteria that will protect all species. In a recent review, Real et al. [31] scanned a database of over 1,000 references on radiation effects in plants and animals. They concluded that no effects could be observed in plants, fish, or mammals for dose rates less than 0.1 mGy h^{-1} (1 Gy $year^{-1}$), but that effects were becoming clear for lifetime exposures at dose rates above 1 mGy h^{-1} (10 Gy $year^{-1}$). UNSCEAR [32] concluded that irradiation at chronic dose rates of 1 mGy day^{-1} to even the most radiosensitive species does not appear likely to cause observable changes in terrestrial animal populations.

Historically, most of the ^{137}Cs doses in caribou and reindeer have been well below these threshold values (Table 5). However, doses in Scandinavian reindeer did rise to several hundred mGy $year^{-1}$ after the Chernobyl accident and were approaching threshold values.

These results should be evaluated in light of the doses from the natural radionuclides ^{210}Pb and ^{210}Po, which can reach 18 mGy $year^{-1}$ in bone and 7–10 mGy $year^{-1}$ in kidney and liver. Although well below the thresholds noted above, most of this dose is due to alpha radiation from ^{210}Po, known to be more damaging than beta or gamma radiation for the same dose. The choice of an alpha radiation weighting factor for nonhuman biota has been the topic of much discussion. The value of 20 used in human dosimetry is probably too high, since it was designed to protect against stochastic effects such as cancer. Chambers et al. [33] recommend an alpha weighting factor of 5 for nonhuman biota, with a range of 1–10. Thomas et al. measured the relative biological effectiveness of ^{210}Po alpha radiation in comparison to 250 kVp X-rays in bovine [34] and porcine [35] endothelial cells, and found values up to 10 for cell survival and clonogenic assay. These are large mammals and may exhibit responses to radiation that are similar to caribou. If a weighing factor of 10 were applied to the ^{210}Po doses in Table 5, then the equivalent doses to bone, kidney, and liver from this natural radionuclide could exceed 100 mSv $year^{-1}$. Since caribou have flourished for tens of thousands of years with this background radiation dose, it is unlikely that a similar dose from ^{137}Cs in the short term would have major consequences for the health of the animals.

Possible Effects in Reindeer

It is difficult to establish experimentally whether these borderline radiation doses are having any measurable effect on caribou and reindeer populations. Most of the knowledge on radiation effects in animals is based on laboratory experiments with large numbers of small mammals such as mice and rats. It is not feasible to carry out similar experiments on the numbers of large mammals that would be needed to establish statistically meaningful results. The best that can be done is to examine natural populations that have been exposed to high levels of radioactivity.

There is some evidence of chromosome aberrations in Norwegian reindeer calves exposed to fallout from the Chernobyl accident. In a study of peripheral blood lymphocytes from 24 calves taken from high-fallout areas in Norway and 26 calves from low-fallout areas, Røed et al. [36] found three dicentrics, two rings, and three translocations in calves from the exposed area, but no dicentrics or rings and only one translocation from the control area. Dicentrics and rings are two types of chromosome aberrations most frequently associated with radiation exposure. The authors cautiously conclude "that certain genetic effects have occurred as a result of the Chernobyl accident in Norwegian reindeer in the most contaminated areas."

In a later study, Røed and Jacobsen [37] reported results from 192 Norwegian reindeer in four herds heavily affected by the accident and three herds only lightly affected. They found significant heterogeneity in the distribution of chromosome aberrations between herds, but concluded that the pattern of chromosome aberration frequencies was not related to the variation in radiocesium exposure. They did confirm that within the most contaminated areas, reindeer born in 1986 showed significantly more chromosome aberrations than those born both before and after 1986.

Impacts on Human Populations

Given the close interrelationship between humans and reindeer/caribou, it is fitting to conclude this chapter with a brief discussion of the impacts on human populations dependent on these animals. Tracy et al. [38] published an assessment of all radiocesium body burden measurements carried out from 1963 to 1990 on indigenous peoples of the Canadian Arctic. The results varied widely with time, geographical location, age, and sex. Average radiocesium concentrations in people for each community were strongly correlated with the concentrations in local caribou herds available in the local community. A comparison of results from the 1960s and 1980s showed that body burdens in people were declining faster than the radiocesium levels in caribou. The authors attributed this rapid decline to a shift from traditional foods to market foods during this period. The cumulative radiocesium doses in people over the entire period varied from 0.3 to 40 mSv (average for each community) with an Arctic-wide average of 12 mSv. It is unlikely that any health effects to humans would be observable at these doses, given that the average worldwide exposure to natural background radiation is about 2 to 3 mSv year^{-1}.

With regard to exposures from ^{210}Po in caribou meat, Thomas [39] estimated average radiation doses of 0.3 mSv year^{-1} and 0.4 mSv year^{-1} to residents in the two communities studied. She obtained the doses to the caribou and wolves to vary from 7 to 20 mSv year^{-1}, assuming a weighting factor of 20 for ^{210}Po alpha radiation.

Because of concerns about heavy Chernobyl fallout over northern Europe, a number of epidemiological studies have carried out on reindeer herders. Kurttio et al. [40] examined the cancer incidence in a cohort of 34,653 people from northern Finland. A total of 1,580 cancer cases were observed versus 1,948 expected on the basis of incidence rates in Northern Finland [standardized incidence ratio (SIR) = 0.81 with 95% confidence interval (CI) = 0.77–0.85]. No association between the lifetime cumulative radiation exposure from global radioactive fallout and cancer incidence in the Arctic population was found.

With regard to the effects of fallout from atmospheric weapons testing before the Chernobyl accident, Wiklund et al. [41] studied cancer incidence from 1961 to 1984 in a cohort of 2,034 Swedish reindeer breeders and found 100 cancer cases observed versus 163 expected. No increased risk was found for the cancer sites considered to be most sensitive to radiation.

From the above, it appears that no observable health effects are likely to occur in human populations dependent on caribou or reindeer. Perhaps the greatest impacts are economic and cultural. In the year following the Chernobyl accident, 73,000 reindeer had to be destroyed in Sweden alone because of high radiocesium levels. This placed a severe economic burden on the Sami people and on the national governments who had to compensate the Sami for losses.

Although the actual amounts of Chernobyl fallout in Canada were minimal, fears of radioactive contamination likely contributed to the decline in traditional food consumption of many northern Canadians. The replacement of traditional foods by market foods places a greater economic burden on northerners who are eking out a marginal existence at the best of times. Futhermore, the hunting and sharing of caribou meat has played a major role in the culture of Canada's aboriginal peoples for thousands of years. This perspective can best be summed up in the words of an Inuit caribou hunter, Pauloosie Kilabuk, from Nunavut Territory:

> ▸ I don't hunt just for me. I hunt for other people. I go out and get a caribou and feel good about myself. It keeps me close to the men I hunt with. I make my parents, kids, relatives and friends happy because they don't have caribou sometimes, and we all come together and share the meat. Caribou is more important than seal to keep my family and community together. With a caribou you can get four or five families together. What is a community feast without a caribou?

Future Directions

Radiocesium levels became elevated in the environment during the latter half of the twentieth century as a result of atmospheric nuclear weapons testing and the Chernobyl nuclear reactor accident. This fallout has been persistent in northern ecosystems, particularly the lichen → caribou/reindeer → human food chain. Radiocesium is disappearing from this system with an effective halftime of 5–10 years and has now decreased to insignificant levels, expect for occasional pockets of high contamination in some parts of northern Europe. Radiocesium contamination of caribou/reindeer will not be considered a problem in the twenty-first century, unless there is another major nuclear event comparable to Chernobyl. If this should occur, there will be a large body of available knowledge gleaned from the twentieth-century investigations.

High levels of the natural radionuclides ^{210}Pb and ^{210}Po have always been present in caribou/reindeer. The available evidence shows that these levels have been constant, at least for the past century. However, global warming, especially in polar regions, could affect this situation. Melting permafrost could lead to higher radon emanation rates from soil, and this could lead to a buildup of higher ^{210}Pb and ^{210}Po levels in the environment.

These natural radionuclides should continue to be monitored in the environment generally and in the lichen → caribou/reindeer → human food chain particularly as harbingers of climate change.

The question of possible radiation effects in caribou/reindeer cannot be answered unequivocally at present. Recent levels of radioactivity have been just below the known threshold values for possible effects. It is possible that contaminants including radionuclides could interact with other stressors to weaken the survivability of caribou and reindeer in the wild. This complex ecosystem should be the subject of ongoing study.

Bibliography

Primary Literature

1. Flexner SB, Hauck LC (eds) (1987) The Random House Dictionary of the English Language, 2 (unabridged)th edn. Random House, New York, pp 315–316
2. Department of the Environment and Natural Resources, Government of the Northwest Territories (September 2009). www.enr.gov.nt.ca. See also the globe and mail, Wednesday, 5 May 2010, p A4
3. United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) (2004) Exposures of workers and the public from various sources of radiation. Report to the general assembly
4. Cornett J, Tracy B, Kramer G, Whyte J, Moodie G, Auclair JP, Thomson D (2009) Polonium-210: lessons learned from the contamination of individual Canadians. Radiat Prot Dosim 134(3–4):164–166
5. Francis W, Chesters G, Haskin LA (1970) Determination of the ^{210}Pb mean residence time in the atmosphere. Environ Sci Technol 4:586
6. Gorham E (1959) A comparison of lower and higher plants as accumulators of radioactive fallout. Can J Bot 37:327
7. Hofmann W, Attarpour N, Lettner H, Turk R (1993) ^{137}Cs concentrations in lichens before and after the Chernobyl accident. Health Phys 64(1):70–73
8. Hanson W (1982) ^{137}Cs concentrations in northern Alaskan Eskimos, 1962-79: Effects of ecological, cultural, and physical factors. Health Phys 42(4):433–447
9. White RG, Holleman DF, Allaye-Chan AC (1986) Radiocesium concentrations in the lichen-reindeer/caribou food chain: before and after Chernobyl. Rangifer Appendix 1:24–29
10. Hutchison-Benson E, Svoboda J, Taylor HW (1985) The latitudinal inventory of ^{137}Cs in vegetation and topsoil of Northern Canada, 1980. Can J Bot 63:784–791
11. Taylor HW, Hutchison-Benson E, Svoboda J (1985) Search for latitudinal trends in the effective half-life of fallout ^{137}Cs in vegetation of the Canadian Arctic. Can J Bot 63: 792–796
12. Holtzman RB (1966) Natural levels of lead-210, polonium-210 and radium-226 in humans and biota of the Arctic. Nature 210:1094–1097, June 11
13. Beasley TM, Palmer HE (1966) Lead-210 and polonium-210 in biological samples from Alaska. Science 152:1062–1063
14. Kauranen P, Miettinen JK (1969) ^{210}Po and ^{210}Pb in the Arctic food chain and the natural radiation exposure of Lapps. Health Phys 16:287–295
15. Jaworowski Z (1966) Temporal and geographical distribution of radium D (lead-210). Nature 212:886–889
16. Perrson RBR (1973) Stable lead and ^{210}Pb in the food chain lichen-reindeer-man. IAEA-SM-175/18
17. Skuterud L, Gwynn JP, Gaare E, Steinnes E, Hove K (2005) ^{90}Sr, ^{210}Po and ^{210}Pb in lichen and reindeer in Norway. J Environ Radioact 84(3):441–456
18. Hanson WC, Palmer HE (1965) Seasonal cycle of ^{137}Cs in some Alaskan natives and animals. Health Phys 11: 1401–1406
19. Westerlund EA, Berthelsen T, Bertig L (1987) Cesium-137 body burdens in Norwegian Lapps, 1965-1983. Health Phys 52(2):171–177
20. Rissanen K, Rahola T (1990) Radiocesium in lichens and reindeer after the Chernobyl accident. Rangifer (Special issue 3): 55–61
21. Åhman B, Åhman G (1994) Radiocesium in Swedish reindeer after the Chernobyl fallout: seasonal variations and long-term decline. Health Phys 66(5):505–512
22. MacDonald CR, Elkin BT, Tracy BL (2007) Radiocesium in caribou and reindeer in northern Canada, Alaska and Greenland from 1958 to 2000. J Environ Radioact 93:1–25
23. Rissanen K, Rahola T (1996) Radioactivity levels in foodstuffs in Finnish Lapland. In: Proceedings of the 11th Meeting of the Nordic Society for Radiation Protection, Reykjavik
24. Arctic Monitoring and Assessment Programme (1998) AMAP Assessment report: arctic pollution issues, chapter 8. Radioactivity. ISBN 82-7655-061-4
25. Holleman DF, Luick JR, Whicker FW (1971) Transfer of radiocesium from lichen to reindeer. Health Phys 21(5): 657–666

26. Shulert AR (1962) Alaskan Eskimos for whom the caribou is a dietary staple have a high strontium-90 concentration. Science 136(3511):146–148
27. Blanchard RL, Kearney JW (1967) Natural radioactivity and ^{137}Cs in Alaskan caribou and reindeer samples. Environ Sci Technol 1:932–939
28. Blanchard RL, Moore JB (1970) ^{210}Pb and ^{210}Po in tissues of some Alaskan residents as related to consumption of caribou or reindeer meat. Health Phys 18:127–134
29. Macdonald CR, Ewing LL, Elkin BT, Wiewel AM (1996) Regional variation in radionuclide concentrations and radiation dose in caribou (*Rangifer tarandus*) in the Canadian Arctic; 1992-94. Sci Total Environ 182:53–73
30. Thomas P, Sheard JW, Swanson S (1994) Transfer of ^{210}Po and ^{210}Pb through the lichen-caribou-wolf food chain of northern Canada. Health Phys 66(6):666–677
31. Real A, Sundell-Bergman S, Knowles JF, Woodhead DS, Zinger I (2004) Effects of ionising radiation exposure on plants, fish and mammals: relevant data for environmental radiation protection. J Radiol Prot 24:A123–A137
32. United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) (1996) Sources and effects of ionising radiation. Report to the general assembly, Supplement No. 46 (A/51/46), Annex: "Effects of Radiation on the Environment," United Nations Sales No. E96.IX.3)
33. Chambers DB, Osborne RV, Garva AL (2006) Choosing an alpha radiation weighting factor for doses to non-human biota. J Environ Radioact 87(1):1–14
34. Thomas PA, Tracy BL, Ping T, Wickstrom M, Sidhu N, Hiebert L (2003) Relative biological effectiveness (RBE) of ^{210}Po alpha-particles versus X-rays on the lethality of bovine endothelial cells. Int J Radiat Biol 79(2):107–118
35. Thomas PA, Tracy BL, Ping T, Baweja A, Wickstrom M, Sidhu N, Hiebert L (2007) Relative biological effectiveness of alpha radiation in cultured porcine aortic endothelial cells. Int J Radiat Biol 83(3):171–179
36. Røed KH, Eikelmann IMH, Jacobsen M, Pedersen Ø (1991) - Chromosome aberrations in Norwegian reindeer calves exposed to fallout from the Chernobyl accident. Hereditas 115(3):201–206
37. Røed KH, Jacobsen M (1995) Chromosome aberrations in Norwegian reindeer following the chernobyl accident. Mutat Res Lett 346(3):159–165
38. Tracy BL, Kramer GH, Zielinski JM, Jiang H (1997) Radiocesium body burdens in residents of northern Canda from 1963-1990. Health Phys 72(3):431–442
39. Thomas PA (1994) Dosimetry of ^{210}Po in humans, caribou, and wolves in northern Canada. Health Phys 66(6):678–690
40. Kurttio P, Pukkala E, Ilus T, Rahola T, and Auvinen A (2010) Radiation doses from global fallout and cancer incidence among reindeer herders and Sami in northern Finland. In: Proceedings of Third European IRPA Congress, 14–16 June, Helsinki, Finland
41. Wiklund K, Holm LE, Eklund G (1990) Cancer risks In Swedish Lapps who breed reindeer. Am J Epidemiol 132(6):1078–1082

Books and Reviews

Advisory Committee on Radiological Protection (2002) Protection of non-human biota from ionizing radiation. Canadian Nuclear Safety Commission INFO-0730

Arctic Monitoring and Assessment Programme (2002) AMAP Assessment 2002: radioactivity in the arctic. Available on-line from the AMAP web site http://www.amap.no. Accessed July 2010

Indian and Northern Affairs Canada (2003) Canadian arctic contaminants assessment report, ISBN-10-0662334663. Library of Congress TD182.4.N67 C36 2003

Indian and Northern Affairs Canada (2009) Canadian arctic contaminants heath assessment report. Available on-line at http://www.ainc-inac.gc.ca/nth/ct/ncp/pubs/har/har-eng.asp. Accessed July 2010

Muir DCG, Shearer RG, Van Oostdam J, Donaldson SG, Furgal C (2005) Contaminants in Canadian arctic biota and implications for human health: preface. Sci Total Environ 351–352:1–3

Thomas DJ, Tracy B, Marshall H, Norstrom RJ (1992) Arctic terrestrial ecosystem contamination. Sci Total Environ 122:135–164

United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) (2006) Report to the general assembly, vol 1. Effects of ionizing radiation

Williams TM, Heard DC (1986) World status of wild Rangifer tarandus populations. Rangifer 1:19–28

Radiation in the Environment, Sources of

Edward J. Waller
University of Ontario Institute of Technology, Oshawa, ON, Canada

Article Outline

Glossary
Definition of the Subject and Its Importance
Introduction
Sources of Radioactivity in the Environment
Future Directions
Bibliography

Glossary

Anthropogenic Created or generated by human activity.

Atmosphere The term atmosphere derives from the Greek *atmosphaira* (atmos = vapor + sphaira = sphere) and on Earth is defined by seven layers extending from the surface of the Earth as follows: troposphere, tropopause, stratosphere, stratopause, mesosphere, mesopause, and thermosphere. Above the thermosphere is normally termed exoatmospheric.

Atoll String of tightly packed small coral islands that enclose or almost enclose a shallow lagoon.

Biocenosis A group of interacting organisms that live in a particular habitat forming an ecological community.

Critical group The target population for an environmental dose assessment. Generally, the members of a population most sensitive to radiation exposure.

Cosmogenic Pertaining to origin, structure, and dynamic behavior of the universe.

Decay chain A series of radioactive transformations, starting with an initial radioisotope that decays into another radioisotope and so on, until a final stable end isotope is reached.

Environment Components of the Earth, and includes (a) land, water, and air, including all layers of the atmosphere; (b) all organic and inorganic matter and living organisms; and (c) the interacting natural systems that include components referred to in (a) and (b).

Half-life The time it takes for a large quantity of radioactive atoms to fall to half of their initial quantity.

HBRA High background radiation area. A region or locality which has levels of natural background radiation in excess of the current limit for radiation workers of 20 mSv a^{-1}. The highest of these areas may be termed very high background radiation areas (VHBRA) when their estimated potential public exposure may exceed 50 mSv a^{-1}.

HLNRA High levels of natural radiation area. Same as HBRA.

Isotope Atom with the same number of protons (Z) as another atom, but different number of neutrons (N), giving it a different atomic mass (A).

JCO Formerly the Japan Nuclear Fuel Conversion Co., a subsidiary of Sumitomo Metal Mining Co.

NORM Naturally occurring radioactive material.

TENORM Technologically enhanced NORM.

Primordial Existing from the beginning.

Pyrophoric A material which can spontaneously combust in air.

Radioactive Unstable isotope that decays by emitting ionizing radiation.

Radioisotope Radioactive isotope; also termed radionuclide.

Radionuclide Same as radioisotope.

RDD Radiological dispersal device.

RORSAT Radar Ocean Reconnaissance Satellite (USSR).

RTG Radioisotopic thermoelectric generator.

SNAP Systems for Nuclear Auxiliary Power.

Spallation A nuclear reaction in which a high energy particle strikes another particle and causes it to fragment into multiple pieces.

Terrestrial Pertaining to the Earth (or more generally, pertaining to land).

TNT Trinitrotoluene (chemical explosive).

Definition of the Subject and Its Importance

Radioactivity in the environment is ubiquitous. From the first moments of time during formation of the universe, elementary particles (protons, neutrons, and electrons) underwent nuclear reactions to produce elements and isotopes, forming all matter in the universe, on a time frame that extends billions of years. Elementary particles formed into nuclear configurations (atoms), consisting of protons and neutrons in the central core (nucleus) surrounded by electrons. By convention, configurations with the same number of protons in the nucleus are termed elements, which are the building blocks of all matter. All elements have members with varying numbers of neutrons in the nucleus, termed isotopes. Some configurations are stable, meaning that they had achieved an optimum number of protons and neutrons in the nucleus. Some configurations were formed with excess protons or excess neutrons in the nucleus, making them unstable. To become stable, these configurations release energy as particles, energy quanta or both. The release of energy as particles or quanta is termed radiation, and these isotopes are termed radioisotopes (or radionuclides). Thus, in the creation of the universe, as atoms formed into more complex structures, a certain amount of radioactive and nonradioactive material made up the structures. As a result, it may be speculated that most structures in the universe have varying levels of naturally occurring radioactive material, which was incorporated into their formation from the beginning of time. The Earth, for example, is a planet consisting of large quantities of naturally occurring radioactive material.

In the past approximately 100 years, humans have learned to utilize the power of the atom for generation of electrical power, performing medical treatment and diagnosis, application to industrial techniques that are not readily achievable through other means, and for the production of weapons. All of these applications give rise to potential introduction of anthropogenic radioisotopes into the environment. The significance of radioisotopes in the environment is multi-faceted. Since radioisotopes were naturally ubiquitous from the beginning of time, it may be thought that evolution of life on Earth has been influenced, in part, by the natural radiation environment that all living organisms were, and are, exposed continuously. All sources of radiation can potentially produce a radiation dose and subsequent biological effect in both human and nonhuman biota. It is known that extremely large doses of radiation can be biologically harmful to living organisms. At low doses of radiation exposure, there is no consistent body of knowledge to describe the biological significance on living organisms. There exists evidence to suggest that low-level radiation exposure may be beneficial, and there is contradictory evidence to suggest that low-level radiation exposure may be harmful. As a result, it is important to continually assess and understand the sources of radiation that exist in the environment and to estimate exposure to human and nonhuman biota. One fact is certain: There will continue to be natural radiation sources in the environment that expose all living organisms on Earth, and in space radiation environments that will expose all future space explorers.

Introduction

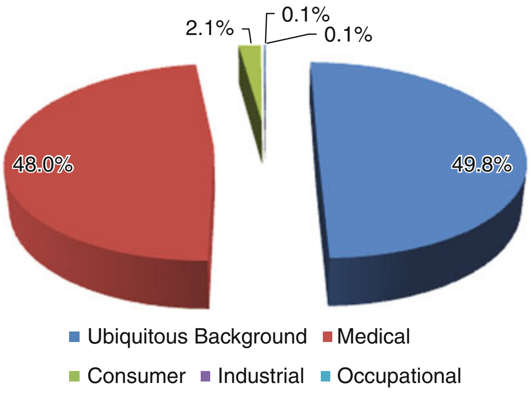
Radioisotopes are fundamentally as old as the universe itself. However, the study of radioisotopes, radiation, and effects on the environment is a little over a century old. As such, environmental radioactivity as a science is a very young field dealing with very old phenomena as the Earth is estimated to be over 4.5 billion years old. It is well known that the Earth is rich in radioisotopes that occur naturally. In addition to the radioisotopes that made up the Earth from primordial formation, natural nuclear reactors existed on the Earth approximately 2 billion years ago, which produced radioactive by-products via nuclear fission of natural uranium. Also, it is also known that radioisotopes are produced on Earth via cosmic ray interactions in the atmosphere. Humans live in a complex radiation environment simply from natural sources, termed naturally occurring radioactive material (NORM). Early investigations in mining materials that contain radioisotopes produced some of the first anthropogenic sources introduced into the environment. Although these sources were natural in origin, their processing and accumulation are such that they do not strictly satisfy the condition of being purely natural, and are often called technologically enhanced naturally occurring radioactive materials (TENORMs). This category of environmental sources can come from many industries such as mining, fertilizer, pulp and paper, and petroleum. Another significant source of radionuclides detected in the environment is from nuclear weapons, specifically historical nuclear weapons testing. Many of the isotopes detectable at low levels in environmental samples are a direct result of nuclear weapons testing, and would not otherwise be present in environmental samples. A routine industrial-origin radionuclide source in the environment is from nuclear reactor operations. Nuclear reactors release low levels of radionuclides into the environment, as allowed by regulatory bodies. Nuclear reactor operations have the distinction of being one of the safest industries, and accidents are rare events. However, nuclear reactor accidents are another potential source of radionuclides in the environment. The only reactor accident that has introduced significant and widespread (compared to natural background) levels of radionuclides to the environment is from the Chernobyl NPP Unit 4 accident in the Ukraine. Commercially produced radioisotopes that may be introduced into the environment comprise the categories of medical and industrial radioisotopes. Medical radioisotopes are routinely released to the environment via nuclear medicine procedures, whereas industrial radioisotopes are typically only released when used in tracer studies. Lost or orphan sources comprise an increasingly more important sector of environmental radioactivity insofar as there are thousands of anthropogenic sources worldwide that are not accounted for that either exist in the environment or may impact the environment. A final category of potential environmental source is that from a radiological dispersal device, or a "dirty bomb." This is a special case insofar as the use of radioisotopes as a weapon suggests targeting of both human and nonhuman biota.

The current breakdown in the USA of collective dose is shown in Fig. 1. This plot is demonstrative of the fact that natural background and medical exposure contribute the majority of dose to humans. It is worthy to note that the actual percentages will vary from country to country on a collective basis. In less technologically advanced countries, the natural radiation component will be much larger than depicted in Fig. 1.

The complex sources of radiation on Earth can be discussed in terms of naturally occurring radiation, radiation originating from nuclear technology origins, and enhanced radiation originating from nonnuclear technologies. The following sections detail the variety of sources of radiation that may have an impact on human and nonhuman biota.

Sources of Radioactivity in the Environment

Of primary importance to the study and early interest in environmental sources of radioactivity were the discovery of x-rays by Wilhelm Röntgen in 1895, radioactivity by Henri Becquerel in 1896, and radium by Marie and Pierre Curie in 1898. The further discovery of nuclear fission by Otto Hahn, Lise Meitner, and Fritz Straßmann in 1938 initiated the atomic age and widespread development of anthropogenic sources of radiation. As such, the study of anthropogenic sources of radioactivity in the environment rapidly became important.



Radiation in the Environment, Sources of. Figure 1 Breakdown of collective dose in the USA (Adapted from NCRP [1])

All anthropogenic radionuclides in the environment have their origins in nuclear technology, through production in nuclear reactors, particle accelerators, mining operations, or nuclear weapon use. Some radionuclides that are natural in origin get concentrated through industrial and chemical processes and are termed "technologically enhanced naturally occurring radioactive material." Radioactivity in the environment can therefore be broken down into two primary categories: (1) naturally occurring and (2) nuclear technology origins. The following sections discuss the various forms of natural and anthropogenic radioactivity and place them into context of sources of radiation in the environment.

Natural Radiation

There are two primary sources of natural radiation on Earth. The first is radioactive material in the Earth itself that originated from planetary formation (terrestrial radiation) and the second is the contribution from radionuclides that have origins from extraterrestrial events (cosmogenic radiation). A third source of environmental radioactivity that is an extension of the cosmogenic component is the radiation environment in aircraft at altitude. Finally, a fourth source, which is important for human space activities, is the natural space radiation environment. Some of the more important radionuclides with respect to human exposure will be discussed, including:

- H-3 (tritium)
- Be-7
- C-14
- Na-22
- K-40
- Rb-87
- Uranium decay series
- Thorium decay series
- Actinium decay series

The order of discussion will begin with the radiation source furthest away from the collective human receptor: the space environment. This outward-in approach is appropriate since (a) all terrestrial radiation originated from events that happened in space, (b) the Earth is completely surrounded by radiation, and (c) over the past 50 years human activities have extended further into space, and the likelihood is high that human space exploration beyond the Earth-moon system will take place within the next 50 years. Following the space environment, the cosmogenic environment, which is the transition between the space and terrestrial environment, will be discussed. Next, a modification to the cosmogenic environment, the high altitude aircraft radiation environment, will be considered as there are implications on flight crew exposure. Lastly, the discussion culminates to consider the terrestrial environment.

Space Radiation Environment By conventional definition of the "environment," exoatmospheric exposures are not typically included. However, as humans (and nonhuman biota) venture further from the immediate vicinity of planet Earth, the effect of the space radiation environment becomes important for exposure profiles, and is therefore discussed herein.

The natural space radiation environment consists of electrons, protons, and heavy ions of various energies, either trapped by the Earth's magnetic field or passing through typical spacecraft trajectories or orbits. The near-earth geospace or magnetosphere is a geomagnetic cavity formed by the Earth's magnetic field as it passes through the solar wind. The cavity is asymmetric, ranging from a 10–12 R_E ("earth-radius," R_E is approximately equal to 6,500 km) hemisphere (sun-side) to a 60 R_E cylinder (night-side). As the Earth penetrates the solar wind, it generates a "wake" or "magnetotail" path in the solar wind for approximately 500 R_E.

The Earth's total magnetic field is the algebraic sum of the Earth's internal magnetic field (molten core and crust residual magnetism) and the net external effects established by solar wind–magnetosphere interactions. The total field is nonconstant, due to slow variations in the internal component, with erratic (random solar flares) and cyclic (solar cycle, seasonal tilt, diurnal effects) variations in the external component.

Although to a first approximation the total magnetic field is dipolar, non-dipolar contributions are important and are best described by numerical models that also account for the offset (approx. 11°) and tilt (approx. 500 km into the Western Pacific) of the geomagnetic axis with respect to the rotational axis. This offset and tilt brings the plasmasphere, also known as the Van Allen radiation belts, to very low altitudes over the coast of Brazil, creating the "South Atlantic Anomaly" (SAA). The SAA is directly responsible for nearly all trapped radiation dose received by space vehicles in low-earth orbits (LEO). Conversely, axial offset and tilt causes the radiation belts to move to higher altitudes (180° longitude from the SAA) over the coasts of Thailand and Vietnam, creating the Southeast Asian Anomaly.

Trapped Radiation Energetic charged particles trapped in the Earth's radiation belts or plasmasphere routinely interact with spacecraft operating in near-earth orbits. The composition, energy, and strength of the trapped particle populations are functions of many variables, the most important being the location at which the radiation is encountered within the trapped radiation belts. At altitudes greater than approximately 80 km, the magnetosphere contains relatively large numbers of trapped electrons and protons, as well as smaller numbers of low-energy heavy ions. The particles interact electromagnetically with the Earth's magnetic field: (1) by gyrating around and traveling along field lines, reflecting between regions of maximum field strength on the lines (conjugate mirror points), and (2) with positively charged protons and heavy ions drifting westward and negatively charged electrons drifting eastward across field lines, around the Earth. Calculation of radiation fluxes integrated over the magnetic field and dipole shell coordinates ("*B–L*" plots) of the orbits provide realistic means of quantification of the trapped radiation component of the natural space environment. The strong dependence of trapped particle fluxes upon particle location is primarily an altitude and latitude dependence. Importantly, specific boundaries exist in radiation trapping regions of the magnetosphere, but these boundaries (particle energy-dependent) are typically smeared areas, quite sensitive to magnetospheric perturbations or variations.

Energetic electrons populating the plasmasphere can be divided into two groups or belts: (1) inner and (2) outer electrons. The inner belt extends to approximately 2.4 R_E on the equator. A region known as the "slot" is found between 2.4 and 2.8 R_E on the equator, wherein the trapped electron, temporally dependent population varies directly with the degree of solar activity. Beyond 10–12 R_E, geomagnetic field intensity has fallen and continuous, steady-state trapping is no longer possible. The intensity and energy bounds of outer belt electrons are greater than those of inner electrons, with peak outer belt fluxes approximately ten times those of inner belt fluxes and with electron outer belt maximum energies approximately two orders of magnitude larger than inner belt energies (10^7 MeV compared with 10^5 MeV). Distant outer belt electrons are brought to low altitude at high latitudes, creating relatively low-intensity and low-energy trapped electron "polar horns."

The proton trapping region of the magnetosphere extends to approximately 4 R_E on the equator. The spatial distribution of these protons varies inversely and monotonically with proton energies, directly opposing trends observed for trapped electrons. Thus, one would expect spacecraft in low-altitude equatorial orbits to encounter the most intense and energetic severe proton fluxes. Unlike lightweight electrons, heavy protons cannot follow the steeply gradated high-latitude geomagnetic field lines, thereby precluding formation of trapped proton polar horns.

The magnetosphere exhibits both short-term and long-term variations in the dynamic behavior of its trapped electron and proton populations. These variations are caused by terrestrial magnetic anomalies, diurnal effects, seasonal tilt, chronology of the 11-year solar (sunspot) cycle and the random occurrences of solar flares.

Transiting Radiation Transiting radiation included in the discussion of the natural space radiation environment has both a solar and an interstellar component. These radiations are termed cosmic rays (CR) and are composed of all the nuclei found in the periodic table (of GeV energies and in various stages of ionization) and energetic (hundreds of MeV) electrons. The distinction between galactic cosmic rays (GCR) and solar cosmic rays (SCR) is made based upon the source and energies of the radiation, rather than its composition.

Galactic cosmic rays are composed mainly of energetic (100–1,000 GeV/nucleon) charged particles such as helium, oxygen, carbon, iron, and nitrogen (and a proportionately low proton and electron component), produced in the interstellar void and arriving omnidirectionally at the surface of the magnetosphere after diffusing through space. Spatial variations in the GCR flux are attributable to variations in the strength and location of the source with respect to the Earth, and the location and depth at which the GCR particles penetrate the magnetosphere. Also, for target locations within 400 km of the Earth's surface, the incident GCR flux is physically shielded on the earth-side by the Earth's shadow.

Solar cosmic rays are energetic charged particles randomly emitted into interplanetary space from events occurring in disturbed regions on, or close to, the surface of the Sun. These particles escape the Sun's near-gravitational field and diffuse throughout the interplanetary medium, traveling along in the solar wind and reaching the magnetosphere within minutes to hours after the solar energetic particle (SEP) event. Solar energetic particle events usually occur in conjunction with solar flares and are composed primarily of 90% energetic protons, 3% alpha particles, and 2% energetic heavy ions and electrons.

Secondary Ionizing Radiation Energetic protons and/or electrons may interact with spacecraft structural materials and cause secondary emissions of an ionizing nature. Proton nuclear capture interactions followed by transmutation and subsequent emission of X and gamma rays can cause ionization in other space-borne materials. Spacecraft design can be used to minimize this secondary radiation hazard. Electron interactions with spacecraft materials can cause more serious problems. For electron energies encountered in the magnetosphere essentially all incident electrons are attenuated by spacecraft walls and shielding layers, ionizing materials and/or causing atomic electron (secondary) excitation. Photo-de-energization of these excited states occurs, with photons possessing a continuous X-ray spectrum emitted approximately along the direction of electron incidence. These bremsstrahlung fluxes typically average one-third that of the incident electron energies, varying linearly with the host material atomic number and with the square of the incident electron energy.

Cosmogenic Environment The space radiation environment consists of energetic charged particles (atomic nuclei, protons, alpha and beta particles) as well as high-energy photons, originating from events from the Sun and beyond the solar system. After entering the Earth's atmosphere, cosmic rays collide with atoms in the atmosphere, resulting in secondary radiation. For example, the interaction of cosmic rays with oxygen and nitrogen results in the production of numerous cosmogenic radionuclides, such as C-14. Cosmogenic radionuclides are therefore those produced continuously in the upper atmosphere via cosmic ray interactions.

The composition and reactions of cosmic rays interacting with the Earth's atmosphere is complex; however, it is dominated by proton and He nuclei interactions (Table 1).

The composition of the atmosphere at sea level is approximately 78% N, 21% O, and 1% other constituents, which is fairly uniform up to approximately 100 km altitude. Therefore, cosmic ray interactions with nitrogen and oxygen are most probable. The high-energy interactions in the upper atmosphere tend to be more exotic than the reactions typically considered on Earth. For example, protons can interact with oxygen or nitrogen via a process called spallation to generate neutrons, protons, and pions. The neutrons and protons can generate further nuclear reactions, generating the cosmogenic radionuclides. For example, a very important reaction involves production of C-14 as follows:

$$ {}_{0}^{1}n + {}_{7}^{14}N \rightarrow {}_{6}^{14}C + {}_{1}^{1}H $$

Radiation in the Environment, Sources of. Table 1
Galactic radiation abundance entering atmosphere

Particle	Mean abundance (%)
Protons	87
He Nuclei	11
Nuclei with $Z > 2$	1
Electrons	1

Source: Adapted from [2].

The importance of C-14 is that it is ubiquitous in the environment and it has a long half-life. This makes it important from both a biological dose standpoint and the fact that C-14 is utilized in carbon dating of organic materials. Carbon dating is possible because while flora and fauna are alive, there is a constant C-14 load within the organism. Upon death, no further C-14 is taken in, and therefore the elimination of carbon is due solely to radioactive decay. Therefore, the age of an organism can be estimated by the C-14 signature.

The principal radionuclides produced via cosmic ray interactions, with half-lives greater than a few hours and sorted by half-life, are presented in Table 2.

The radionuclides produced via cosmic ray interactions eventually settle on Earth through atmospheric dispersion and deposition, as well as through precipitation processes, and become the cosmogenic component of terrestrial radiation.

Aircraft Radiation Environment The Earth is surrounded and continually bombarded by radiation. The extraterrestrial environment has impact both on manned space missions, for generation of cosmogenic radionuclides, and for contribution of human exposure

Radiation in the Environment, Sources of. Table 2
Principal cosmogenic radionuclides

Radionuclide	Half-life	Decay mode
^{10}Be	1.6 million years	β
^{26}Al	0.72 million years	β^{+}
^{36}Cl	0.3 million years	β
^{14}C	5,730 years	β
^{32}Si	330 years	β
^{39}Ar	269 years	β
^{3}H	12.3 years	β
^{22}Na	2.6 years	β^{+}
^{35}S	87 days	β
^{7}Be	53 days	EC
^{33}P	25 days	β
^{32}P	14.3 days	β
^{28}Mg	20.9 h	β
^{24}Na	15 h	β

at aircraft altitude. The exposure due to cosmic radiation is highly variable; however, it may be generally said to increase as both a function of altitude above the Earth's surface due to decreased atmospheric shielding, and increase in latitude from the equator, due to decreased geomagnetic shielding [3].

The dose rate from cosmic radiation increases with depth into the atmosphere to a maximum value (known as the Pfotzer maximum) at approximately 20 km (60,000 ft), whereupon it decreases with decreasing altitude to the Earth's surface [4]. The principal constituents of the dose at aircraft altitudes are neutrons (55%), electrons/positrons (20%), protons (15%), energetic photons (5%), and muons (5%), whereas at surface level almost all dose from cosmic ray origins is attributable to muons.

At altitudes from 3 to 25 km (10,000–80,000 ft), neutrons are the dominant component to the dose [5]. These neutrons are produced through complex cascades involving spallation processes induced by primary cosmic particles (principally protons) entering the atmosphere [6]. The dose to humans in aircraft from the primary neutron spectrum is complicated by interactions of neutrons and other particles within the aircraft aluminum structure and other materials within the aircraft.

Terrestrial Environment The natural terrestrial radiation environment has two principal components: (1) primordial and (2) cosmogenic (discussed in section on "Cosmogenic Environment"). Primordial radionuclides have, by definition, been present since Earth, as a terrestrial body, formed. During the Earth's formation approximately 4.5 billion years ago, a great number of isotopes were introduced as part of the mass generation. Only radioisotopes with long half-lives persist in the Earth to this day. There are a number of long-lived radionuclides that make up the primordial family, presented in Table 3. As may be seen, most of the primordial radionuclides all have half-lives much greater than the estimated age of the Earth.

Besides the single radionuclides presented in Table 3, there are also three principal and very important radionuclide chains that are primordial:

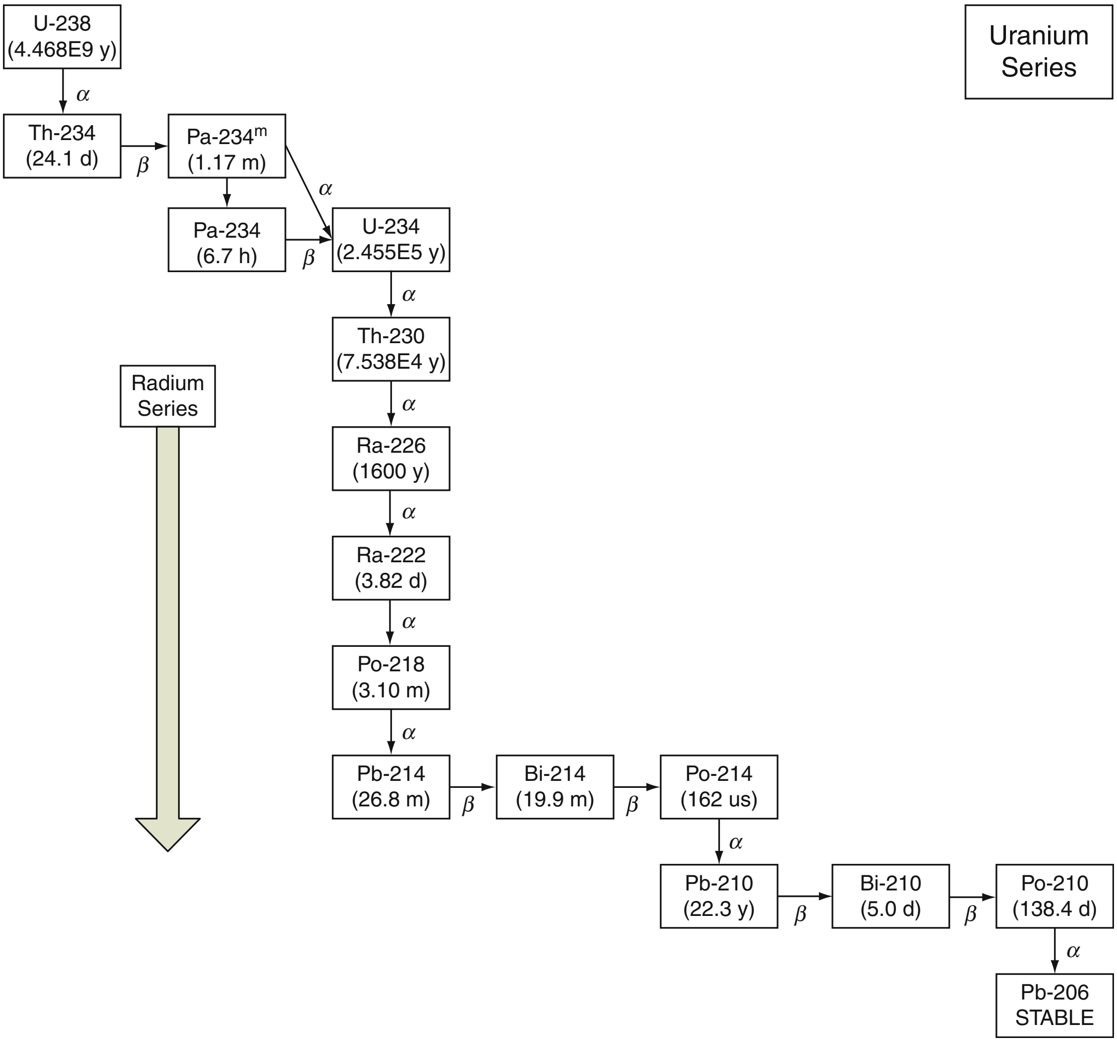
1. U-238 decay chain (uranium series)
2. Th-232 decay chain (thorium series)
3. U-235 decay chain (actinium series)

Radiation in the Environment, Sources of. Table 3
Primordial radionuclides and half-lives

Nuclide	Half-life (billion years)	Decay mode
^{40}K	1.3	β
^{50}V	600,000	β
^{87}Rb	47	β
^{113}Cd	9,000,000	β
^{115}In	460,000	B
^{123}Te	12,000	EC
^{138}La	110	β
^{142}Ce	50,000,000	α
^{144}Nd	2,380,000	α
^{147}Sm	110	α
^{148}Sm	7,000,000	α
^{152}Gd	110,000	α
^{156}Dy	200,000	α
^{176}Lu	27	β
^{174}Hf	2,000,000	α
^{180}Ta	16,000	β
^{187}Re	50	β
^{190}Pt	700	α
^{204}Pb	140,000,000	α
^{238}U	4.5	α
^{235}U	0.7	α
^{232}Th	14.0	α

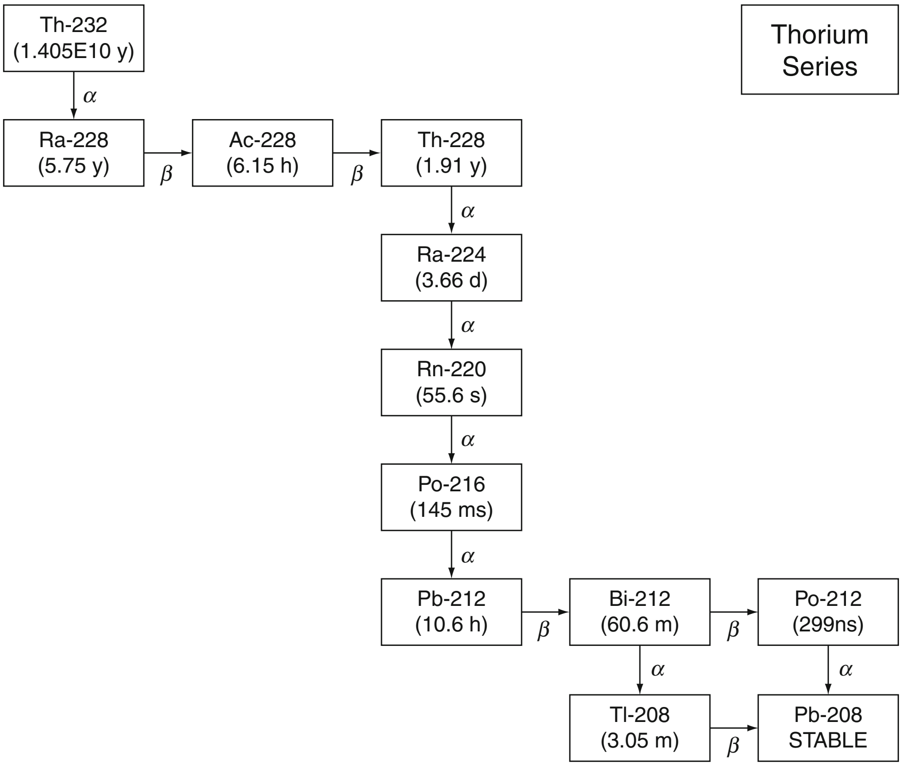
The natural decay series are presented in Figs. 2–4. In the figures, parent radionuclides that undergo alpha decay are above their progeny (sometimes termed "daughter" products), while parents that undergo beta decay are to the left of their progeny. Under each radionuclide identifier, the radiological half-life is presented in brackets. All decay series are seen to eventually decay to stable lead isotopes.

The most important radionuclides from an environmental exposure perspective are K-40 and the uranium and thorium series decay chains. The actinium decay chain is less important due to its lower abundance compared to the other chains. K-40 is ubiquitous in the environment, and is an important radionuclide



Radiation in the Environment, Sources of. Figure 2
Uranium decay series

in humans, with an average activity concentration in the body of approximately 60 Bq kg^{-1} [7]. Humans have a continual uptake of K-40 through ingestion of food, as the average quantity of natural potassium consumed is approximately 2.5 g day^{-1}, where K-40 makes up 0.017% of naturally occurring potassium. From an environmental monitoring standpoint, K-40 is important as an indicator radionuclide. That is, a primordial radionuclide that is almost always expected to be detected in environmental samples. The uranium and thorium series are also ubiquitous and very important to terrestrial radiation due to the progeny. Of greatest importance from the perspective of dose to humans is the radon component commonly termed "radon gas" or "radon" (Rn-222) in the uranium decay chain. There is also a radon component in the thorium series, commonly termed "thoron gas" or "thoron" (Rn-220), but is less significant for dose to humans due to its short half-life of under 1 min. The decay chains are interesting in the fact that the top of the chain is matter in solid form and all progeny are solid matter until radon, which is a gas, and decays back to solid form progeny until the series finally reach stable lead isotopes. This gas form is the reason for the

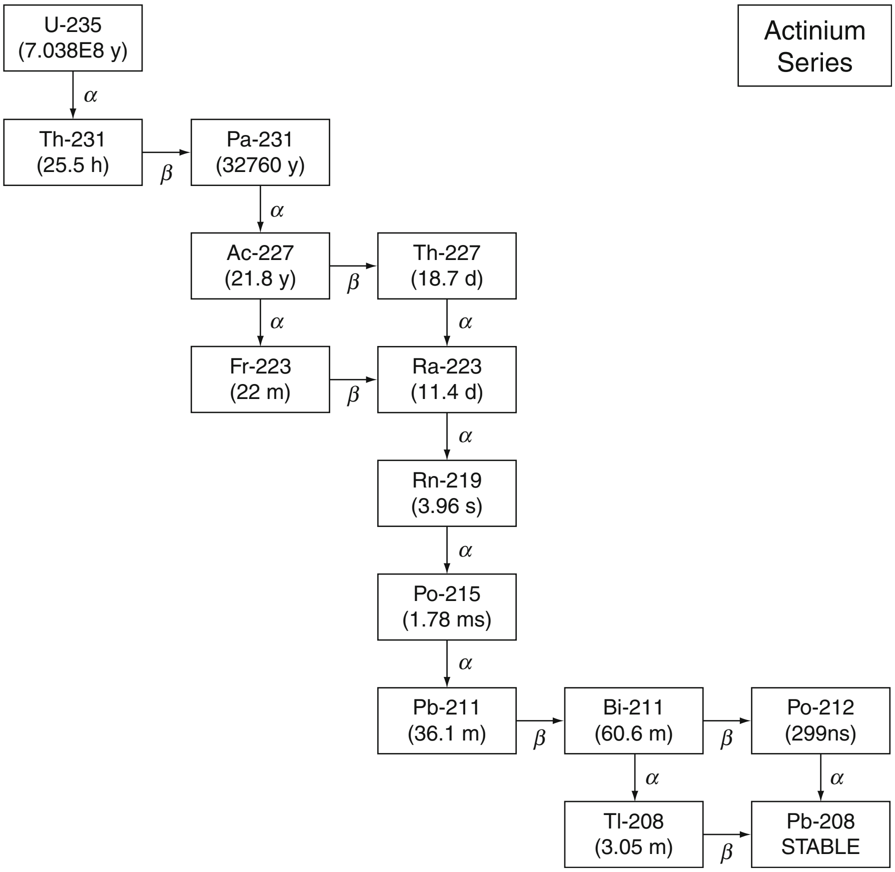


Radiation in the Environment, Sources of. Figure 3
Thorium decay series

high human dose, as it is available for inhalation and instillation in the respiratory tract where it will decay via alpha emission and generate alpha-, beta-, and gamma-emitting progeny. Also, natural series are important as sources of environmental radiation through mining and industrial use of some of the isotopes (e.g., Ra-226 used in luminescent materials, Th-232 used in lantern mantels, and U-235/238 used for nuclear fuel). Some of these uses will be discussed in subsequent sections.

Natural Reactors In 1972, a French physicist named H. Bouzigues made an unexpected discovery during routine measurement of uranium concentrations in the fuel supplied to French nuclear reactors obtained from Oklo in Gabon, West Africa. These concentrations are basically a constant at any given time, changing only through their characteristic half-lives. The U-235 concentration in 1972 was 0.7202 ± 0.00004% and Perrin measured 0.7171%, the first measurement ever recorded outside this range. Other discrepancies were to follow, with some uranium samples depleted to half their normal concentrations. In addition, the presence of some trace levels of fission products was also observed [8]. The observations were explained by a theory that a number of coincidental physical factors merged at Oklo approximately 2 billion years ago to create a natural fission reactor. In fact, in the areas of Oklo and nearby Bangombé, at least 17 uranium deposits are hypothesized to have acted as natural fission reactors [9]. There were a number of prerequisite conditions required to make a natural fission reactor, stated as follows:

1. Sufficient U-235 – At the time of the natural reactors (2 billion years ago), the amount of U-235 in natural uranium was approximately 3% (as opposed to the 0.72% currently). At 3% enrichment, a chain reaction is sustainable (in fact, this is the approximate enrichment in commercial light water–moderated reactors).



Radiation in the Environment, Sources of. Figure 4
Actinium decay series

2. Sufficient physical size – If the "reactor" was too small, then too many fission-produced neutrons would escape, and a chain reaction would not be possible. One of the most studied natural reactors around Oklo is of dimensions 12 m × 18 m × 0.5 m [9].
3. Sufficient moderator – Fast neutrons produced from fission must slow down by colliding with other nuclei. An effective moderator is water, and therefore if the uranium (uranium dioxide, or UO_2) had formed into a porous matrix, then water could impregnate the pores and provide the moderator required to sustain the chain reaction.

Thus it is believed that over 2 billion years ago, natural uranium deposits formed into porous matrices of sufficient size to allow critical reactors to form in this region. Based upon observations of fission product concentrations (especially Xe and Kr) in surrounding mineral deposits, it has been estimated that the reactors ran with 30 min active pulses, separated by 2.5 h dormant periods, continuously cycling for approximately 150,000 years [10]. This cycling occurred when (a) water entered the pores and a chain reaction started, resulting in (b) energy liberated through the chain reaction heating the surroundings and evaporating the water or turning it to steam, thus stopping the chain reaction, until (c) the surroundings cooled and more water could enter, thereby restarting the chain reaction [9].

There is modern interest in studying these ancient natural reactors as they provide a naturally aged laboratory in which to study migration of fission products through the environment, which is a very important area of study in radioactive waste management [11].

Radon Radon was discovered in the early 1900s in Germany by Friedrich Ernst Dorn. Radon is a ubiquitous, odorless, colorless, tasteless, and dense noble gas. Radon is heavier than air, and it is soluble in water. Because it is a noble gas, radon is chemically inert and electrically uncharged. It is a naturally occurring radioactive material which originates from the U-238 decay chain. When uranium, thorium, or actinium atoms decay, other radioactive atoms are formed. One of these is radium. Radium is the immediate source of radon. The radium isotope of principal interest is Ra-226 which decays through alpha emission to produce radon gas (Rn-222). There are 33 isotopes of radon listed in the chart of the nuclides [12], 3 of which are naturally occurring. All of the three emit alpha particles when they decay, and they all have short half-lives. Rn-222 is commonly known as “radon” and is the most stable isotope with a 3.8 day half-life. Rn-220 (or “thoron” as it is referred to in historical context) is a member of the thorium decay series and has a much shorter half-life (55.6 s). Rn-219 (historically referred to as “actinon”) is a member of the actinium series and has the shortest half-life of the three isotopes at 4.0 s. Because it is naturally occurring, radon is found almost everywhere, including air, soil, groundwater, granite, pumice, clay, brick, and other construction materials such as concrete made from fly ash and industrial slag.

The primary human health hazard from radon is indoor exposure. Radon can enter into buildings by several pathways. The surrounding composition of soil and concentration of uranium is the basis for varying levels of radon from location to location. Indoors, the basement is the cause for concern where radon can diffuse through the foundation and travel through cracks, remaining for longer periods of time due to lack of ventilation and minimal exit paths. The building structure plays an important role as mentioned for cracks in the foundation along with wall joints which present a likely entry path for the gas. Cracks in the landscape surrounding a home can also accelerate the entrance of radon into the atmosphere and home.

The World Health Organization (WHO) has estimated average indoor radon concentrations for a number of countries, presented in Table 4. It may be seen that there is significant variation in the average radon concentrations globally.

Radiation in the Environment, Sources of. Table 4
Average global radon concentrations

Country	Average concentration (Bq m^{-3})
Canada	11.4
Czech Republic	140
Finland	123
Germany	50
Ireland	60
Japan	20
Lithuania	37
Norway	51–60
Russia	19–250
Sweden	108
Switzerland	70
UK	20
USA	46

Source: Adapted from [13].

The effective dose calculated for a radon exposure of 200 Bq m^{-3} is 3 mSv a^{-1}, whereas at 600 Bq m^{-3}, the effective dose is 10 mSv a^{-1} [14]. The International Commission on Radiological Protection recommends action levels be established between 200 and 600 Bq m^{-3}, where an action level would indicate a point where remedial action should be taken to reduce the radon concentration.

For nonhuman biota, outdoor concentrations of radon are important as these organisms spend most of their lives in more direct contact with the source of radon gas (i.e., uranium and thorium bearing ores). Although radon gas disperses rapidly in open-air environments [7], closed environments, such as for burrowing animals, may yield elevated radon concentrations.

Considering environmental pathways, radon (Rn-220/222) is the most important terrestrial radionuclide, whereas polonium (Po-210) is the most important radionuclide in aquatic environments. From a yearly dose standpoint, radon gas is the most important environmental radionuclide for human exposure.

High Background Radiation Areas Regions on Earth that have elevated levels of natural radiation compared to the global average may be termed high background radiation areas (HBRA), very high background radiation areas (VHBRA) or high levels of natural radiation areas (HLNRA). The levels of natural radiation can vary greatly depending on location, with the annual effective radiation dose from natural radiation for the world population being approximately 2.4 mSv. Some prominent and inhabited areas with high levels of natural radiation (leading to average local doses much higher than the world average) are found in the following locations (approximate population in brackets):

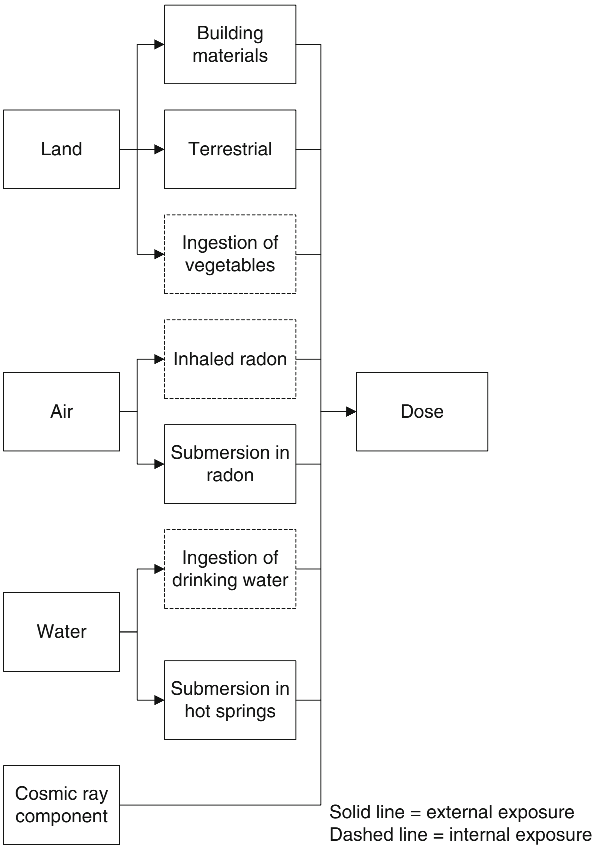
- Ramsar, Iran (2,000)
- Kerala, India (100,000)
- Tamil Nadu, India (30,000,000)
- Mombasa, Kenya (500,000)
- Guarapari, Brazil (100,000)
- Morro do Ferro, Brazil (500)
- Yangjiang, China (3,000,000)

Focus: Ramsar, Iran

Ramsar, a town of 2,000 people, has a long history of being a vacation destination, with approximately 50 hot springs, 9 of which have for centuries been used as recreation and health facilities by the local population and visitors alike. The soil in Ramsar has elevated levels of Ra-226, compared with world average expected values, and the hot springs bring terrestrial radium to the surface, where it precipitates. The radiation fields in Ramsar are heterogeneous with hotspots having dose rates as high as 100 μGy h^{-1}. In some areas of Ramsar, people receive an effective radiation dose of hundreds of mSv a^{-1}, which tends upward of 100 times the average global value [15].

There are two significant pathways for human exposure in Ramsar: building materials and hot springs, both primarily due to radium dissolved in mineral water. Since the raw materials utilized to build the structures come directly from the soil, they are concentrated in Ra-226. In addition, local produce contributes to the effective dose through plant sequestering of Ra-226 and progeny. The typical pathway of exposure from environmental radiation in Ramsar, Iran, is depicted in Fig. 5. It may be seen that the pathways for exposure are varied and complex.

The average and maximum dose rates from various areas around the world are presented in Table 5 (adapted from [16]).



Radiation in the Environment, Sources of. Figure 5
Typical exposure pathways for a resident of Ramsar, Iran

It may be seen that the dose received by persons living in very high background areas can be as high as 100 times the global average dose. Of interest is the fact that in the very high background radiation areas such as Ramsar, no indication of excess cancer in the population has been observed [15].

Naturally Occurring Radioactive Material (NORM) and Technologically Enhanced NORM (TENORM)

Naturally occurring radioactive material (NORM) is germane to the discussion of sources of radioactivity in the environment. NORM, as a source of environmental radioactivity, has been discussed in the section on "Natural Radiation", and does not include anthropogenic radionuclides. It is difficult to separate discussions on NORM from other categories of environmental radioactivity, as there is significant overlap. The fact

Radiation in the Environment, Sources of. Table 5
Average and maximum dose rates for comparison

Location	Average (mSv a^{-1})	Maximum (mSv a^{-1})
High Background		
Ramsar (Iran)	10.2	260
Guarapari (Brazil)	5.5	35
Kerala (India)	3.8	35
Yangjiang (China)	3.5	5.4
Normal Background		
Norway	0.63	10.5
China	0.54	3.0
India	0.48	9.6
Germany	0.48	3.8
Japan	0.43	1.26
USA	0.40	0.88
Denmark	0.33	0.45

that NORM has natural origins does not suggest that NORM cannot have high levels of radioactivity. In fact, through technological processes (technologically enhanced NORM, or TENORM), naturally occurring radionuclides (uranium and thorium decay series, as well as other natural radionuclides, such as K-40) can concentrate to yield reasonably large specific activities in materials, and through recycling of materials NORM can become a significant source of exposure.

NORM is defined as [17]:

- Any of the primordial radionuclides or radioactivity present in soils, rocks, and materials undisturbed as a result of human activities

In contrast, the definition of TENORM is somewhat more complicated [17]:

- Naturally occurring radioactive material disturbed or altered from natural settings or present in a technologically enhanced state due to past or present human activities and practices, which may result in a relative increase in radionuclide concentrations, radiation exposures and risks to the public, and threat to the accessible environment above background radiation levels. "Technologically enhanced" means that the radiological, physical, and chemical properties of the radioactive material have been altered such that there exists a potential for: (a) redistribution and contamination of environmental media (soil, water, air, and biota), (b) increased environmental mobility in soils, surface water and ground water, (c) incorporation of elevated levels of radioactivity or increased accessibility in products and construction material, or (d) improper disposal or use of disposal methods that may result in unnecessary or elevated exposures to individuals and populations via the accessible environment.

Accountability for naturally occurring radioactive material (NORM) and TENORM (technologically enhanced NORM) has been coming under increased scrutiny since the publication of the 1990 International Commission on Radiological Protection recommendations [18] for a public dose limit reduction from 5 to 1 mSv a^{-1}. There are numerous sources of TENORM that are generated through industrial processes and may become available for recycling (Table 6), thereby placing the exposure closer to environmental and human receptors [19].

Although, by definition, NORM is "natural," this does not belay the fact that in a large enough quantity it may become hazardous (the toxicology principle that "the dose makes the poison"). For NORM to be considered hazardous in the recycling industry, there must be a pathway from the NORM to humans and/or the environment from recycling.

Recycled landfill material that is rich in uranium and thorium is one pathway for NORM or TENORM to contribute to public dose. Another significant pathway is through metal recycling, when the metals have been in contact with NORM, usually for significant periods of time (years), such that small quantities of radioactive material have been allowed to collect on surfaces (such as scaling or sludge). It is possible that recycling metal with this type of NORM will result in contaminated metal, or at the very least, contaminated equipment. Another important issue regarding steel recycling is the possibility of having an anthropogenic source (e.g., from a nuclear gauge) mixed in with a shipment containing NORM. If the recycler regards an indication of radiation from this type of shipment as

Radiation in the Environment, Sources of. Table 6
Industrial activities that generate TENORM

Industrial activity	Recycled forms
Uranium mining	Tailing residue
Phosphate fertilizer	Scale, slag
Coal combustion	Fly ash, bottom ash, slag
Oil and natural gas production	Scale, sludge, contaminated water
Municipal water treatment	Sludge, radium-selective resins
Metal mining and processing	Slag, leachate, tailings
Geothermal energy production	Scale, brine-pond deposits
Pulp and paper production	Scale
Metal casting	Ceramic shells
Former radium processing facilities	Contaminated soils

being caused by the NORM, and determines to recycle the scrap, they may inadvertently smelt the gauge source and contaminate all the recycled material.

The risk from recycling NORM (e.g., scaling on steel pipes) would generate insignificant and nonmeasurable contamination due to dilution, and ideally portal monitors and other recycling plant–detection equipment would ignore NORM altogether. However, separating the detection of NORM from non-NORM would be very challenging in the recycling environment, and would likely require a radiation specialist on site. Many recycling companies have radiation detection equipment installed at entry points and/or various locations in the plant. In general, most recyclers have a "zero tolerance" policy for accepting radioactive material. The policy is based upon the alarm indication as provided by the radiation detection equipment, which in turn depend upon the discrimination and sensitivity of the equipment. Therefore, if NORM alarms the radiation detection equipment, this is reason enough to refuse acceptance of a shipment. When a shipment is refused, some sort of action must be taken. The action will clearly depend upon the individual circumstances. The recycling company often initiates the NORM (re)action based upon "blind" instrument alarms, and has very little culpability with respect to this action. In fact, the recycling company has no requirement to accept any given load of material. Litigation and economic penalties surrounding the transport and disposal of materials containing NORM or TENORM are of great concern to many industries that have not previously needed to address the NORM issue [20].

Tobacco Product Tobacco leaf, by virtue of the fact that it is manufactured into smoking product, deserves special consideration. Its categorization here as a technologically enhanced NORM resides in the fact that production of tobacco products is a technological activity, yet the radioactive material is sequestered naturally by the tobacco leaf. Natural radioactivity, primarily Polonium-210 (Po-210) and Lead-210 (Pb-210), are sequestered by tobacco plants from two principal pathways: (1) through the root system by absorption through the ground, or (2) by leaf deposition from the atmosphere. These naturally occurring radioisotopes originate from the terrestrial uranium decay chain found in soil. Regarding environmental radionuclides and impact to human health, Po-210 is one of the most radiotoxic nuclides to humans [21] and is volatile at the temperature of burning tobacco (500–700°C), making it highly available for transport and deposition throughout the human respiratory tract via smoking. The main source of Po-210 and Pb-210 in the atmosphere, which subsequently deposits on tobacco leaves, is Rn-222. The radon gas originates as the progeny of Ra-226 which is present in soil. A second pathway for Po-210 and Pb-210 to enter the plants is via the tobacco plant root systems, again from the uranium decay chain. The concentration of Po-210 and Pb-210 in the soil depends on the humus content of the soil [22]. Soil with higher humus content will generate a higher concentration of these radionuclides in the tobacco plant. The concentration of these radionuclides in tobacco leaves depends on many factors, including the type of soil the plant is grown in, the weather conditions during the time the tobacco is growing and the variety of the tobacco and the storage methods. The importance of radionuclides in tobacco can be illustrated as follows. Depending upon the method of smoking and filtration used, the tobacco used may contain tens of mBq of Po-210 which is vaporized, inhaled, and deposited along the tracheobronchial lining in the smoker, where it can irradiate the sensitive epithelial cells [23]. If a person smokes a single pack (20 cigarettes) per day for 50 years, they will have inhaled from 365,000 cigarettes. Assuming 10 mBq Po-210 per cigarette, and 100% inhalation efficiency (i.e., no effect of filter and no loss of Po-210) the total amount of Po-210 inhaled over 50 years will be 3.65 kBq. This relatively large quantity of Po-210 to the lungs would be expected to increase the risk of lung cancer. Even when the smoking product is filtered, a significant portion of radionuclides are still exposed to the respiratory tract. The risk scales up or down as a function of activity per unit tobacco product, amount smoked, and duration of smoking.

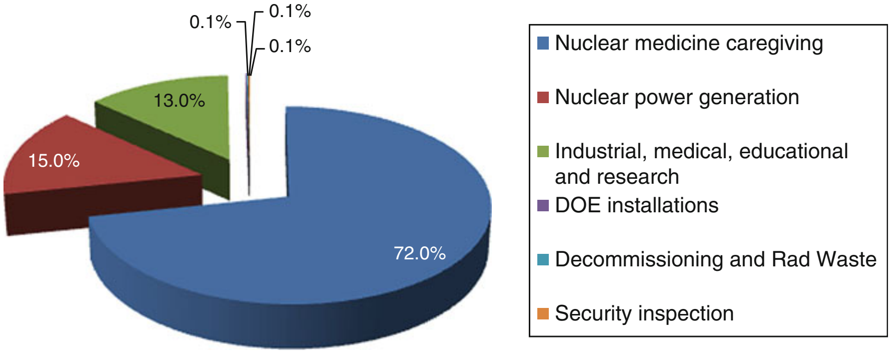
From an epidemiological perspective, there have been no conclusive correlations that tobacco radioisotopes are a major contributor to lung and esophagus cancer from smoking, although it has been well proven that smoking increases radiation dose to the people who smoke, or are exposed to "second-hand" smoke. However, it is well known that smoking is related strongly to lung cancer and therefore it is highly probable that this technologically enhanced environmental radioactivity, which is concentrated in the lungs through smoking, is one of the contributors (amongst a number of other nonradioactive contaminants found in cigarette tobacco) to excess cancer risk observed in the smoking population and likely in the "second-hand smoke" cohort [24]. Infants and children are more radiosensitive to tobacco smoke, as well as persons with weakened immune systems.

Nuclear Reactor Operations

There are three principal components to nuclear reactor operations that can contribute sources of radionuclides to the environment: (1) releases from mining, milling, processing/reprocessing operations; (2) operational releases; and (3) waste storage. Currently, radioactive waste is not a large contributor to sources of environmental radioactivity. However, due to the fact that waste storage solutions are actively being pursued worldwide, this category deserves discussion. Accidental releases related to nuclear power generation are discussed in section on "Nuclear Material Accidents".

It is worthy to note that commercial nuclear power contributes to a very small proportion of the collective effective dose to the population. For example, Fig. 6 shows the breakdown of collective dose to the US population from industrial exposure [1]. It is noted that nuclear power generation is responsible for 15% of the total industrial exposure, where the average effective dose for the exposed population ranges from 1 to 10 μSv, which is, at maximum, approximately 0.05% of the total yearly exposure from natural background radiation.

The nuclear generation component can be further broken down into constituents, as presented in Fig. 7 [1]. It may be seen that uranium mining and milling are the largest contributors (87.2%) to the collective dose



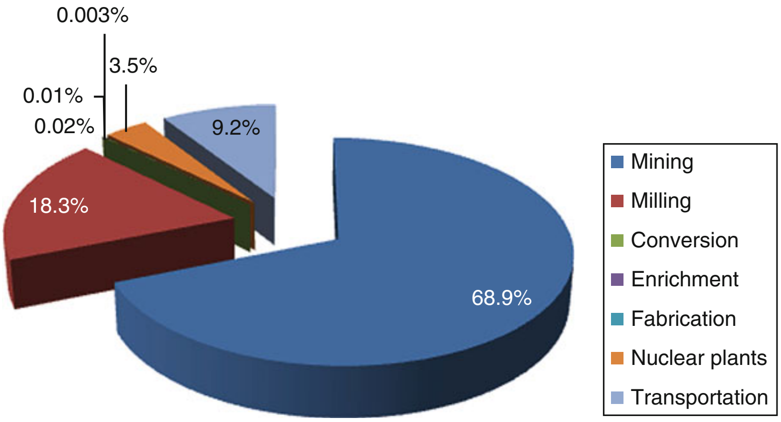
Radiation in the Environment, Sources of. Figure 6
Industrial exposure breakdown of the US population (Adapted from [1])

of the US population, whereas operation of nuclear plants is responsible for only 3.5%.

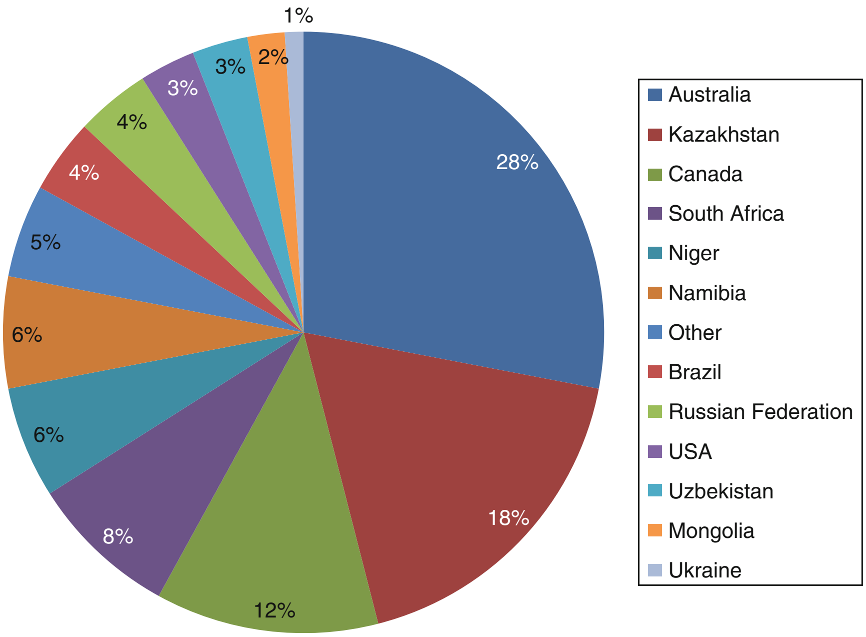
The following section discusses the three principal contributors to environmental sources of radioactivity from nuclear power generation.

Mining, Milling, and Processing of Uranium The approximate geographic breakdown of the world's supply of natural uranium is depicted in Fig. 8.

The life cycle of uranium for use in nuclear power generation slightly differs depending upon whether



Radiation in the Environment, Sources of. Figure 7
Breakdown of nuclear power collective dose (Adapted from [1])



Radiation in the Environment, Sources of. Figure 8
World's supply of uranium

nuclear fuel is to be enriched, and whether fuel reprocessing activities will take place. The generic steps in the nuclear fuel cycle are provided in Table 7.

To start the cycle, an adequate location for uranium mining must be identified. Mining operations generate dust which includes uranium and progeny, and radon is present in large concentrations in the mining environment. Both are contributors to human exposure in and around the mining site. An alternate method of extracting uranium from ore deposits is by in situ leaching (ISL), which is the process of recovering uranium by dissolving it in a leaching agent and extracted via pumps. ISL requires the ore body be permeable to liquids and located such that ground water cannot be contaminated away from the ore body. The main advantage to ISL is that it does not produce tailings. Milling refers to the process of removing unwanted minerals from the ore (including radioactive minerals) to generate the required U_3O_8 product (termed “Yellowcake”). The tailings can contain many undesirable by-products including trace quantities of the mined ore (uranium), other radioactive materials and chemical used in the mining and leaching processes to separate the useful ore. The tailings may be stored above ground, below ground, or in ponds. The potential environmental impact from tailings include (1) radon gas, (2) atmospheric dispersion of tailing material, and (3) potential use of tailing material for building material or landfill [25]. It is worthy to note the fact that tailings have historically been unsecured, suggesting that potential spread of tailing material from the source of generation is not insignificant, and that both human and nonhuman biota are potential receptors. Enrichment and fabrication into fuel both involve chemical processing regardless of whether the end product will be natural or enriched fuel. Chemical processing always produces a waste stream, and therefore environmental spread of uranium and progeny will occur. Significantly, the removal of radium and progeny from the uranium production stream has lead to environmental contamination at some sites [26]. The predominant radionuclides acting as environmental sources of radiation from mining, milling, and processing of reactor fuel are natural uranium isotopes (and progeny), natural thorium isotopes (and progeny) and explicitly Ra-226 and Rn-222, all of which are naturally occurring species.

It must be noted that radionuclides generated as a by-product of mining and milling processes may be considered technologically enhanced naturally occurring radioactive material (TENORM), which was discussed in section on “Naturally Occurring Radioactive Material (NORM) and Technologically Enhanced NORM (TENORM)”.

Radiation in the Environment, Sources of. Table 7
Generic nuclear fuel cycle

Step	Activity	Comment
1	Exploration	Test digs, aerial survey, etc.
2	Mining	Extraction of uranium ore
3	Milling	Yellowcake produced (U_3O_8)
4	Conversion to UO_2 (natural) or UF_6 (for enrichment)	Complex chemical processing
5	*Enrichment (if applicable)*	Increasing U-235 content
6	Fuel fabrication	Into fuel elements or bundles
7	Reactor operation	Depletion of U-235
8	*Reprocessing of spent fuel (if applicable)*	Back to enrichment/fuel fabrication
9	Spent fuel storage	High-level “waste”

Routine Operational Releases Nuclear power plants have multiple barriers to prevent release of radionuclides to the environment. However, under normal operation, routine releases of radionuclides occur as allowed by the regulatory body in the country in which the nuclear power plant is operating. Gaseous forms of radionuclides may vent from locations within the plant. For example, tritium (hydrogen atom) formed in cooling and moderating water can easily move through engineered barriers. Volatile radionuclides that exist in gaseous form such as radioisotopes of xenon, krypton, and iodine can leak from the fuel and ultimately be purged through the filtered air discharge system of the reactor building. The release of radionuclides to the environment is a function of reactor type in use. Data for normalized releases of radionuclides from nuclear power plants are provided

Radiation in the Environment, Sources of. Table 8 Nuclear power plant radionuclide releases

Plant type	TBq GW(e)$^{-1}$ a^{-1}						
	Noble gas	H-3 (gas)	C-14	I-131	Particulate (gas)	H-3 (liquid)	Other liquids
PWR	13	2.4	0.22	0.0002	0.0001	19	0.008
BWR	180	0.86	0.51	0.0003	0.35	0.87	0.01
CANDU	250	330	1.6	0.0001	0.00005	340	0.044
RBMK	460	26	1.3	0.007	0.008	11	0.006
FBR	210	49	0.12	0.0002	0.001	1.7	0.023

in Table 8 (adapted from [8]). The types of reactors considered are PWR (Pressurized Water Reactor), BWR (Boiling Water Reactor), CANDU (CANadian Deuterium Uranium), RBMK (Reaktor Bolshoy Moschnosti Kanalniy *(Russian)*, meaning high power channel-type reactor), and FBR (Fast Breeder Reactor).

As is evidenced by the collective dose component of nuclear power from Fig. 7, radionuclides released from nuclear power generation do not contribute an appreciable exposure.

Nuclear Versus Coal It has long been known that burning coal releases radionuclides to the environment [27]. When comparing the emissions of radioactivity from nuclear power plants with those from coal-fired power plants, it is crucial to understand the radionuclides involved and their source. For coal-fired power plants the source of the radioactivity is the naturally occurring radioactive material within the coal itself. The main radionuclides present include the U-238 series, Th-232 series, and K-40 with typical concentrations of 20, 20 and 50 Bq kg^{-1}, respectively [28]. These values, however, can vary widely depending on the type of coal. During the combustion process a by-product known as fly ash forms. Radionuclides can either be encased within the fly ash or attached to the surface of the fly ash. As coal burns, some of the more volatile radionuclides can escape as gas but most will escape on the fly ash. Activity concentration on the fly ash is about ten times higher than the feed coal. In modern coal-fired power plants, approximately 99.5% of the fly ash is prevented from escaping to the atmosphere (compared to <90% in older plants). Doses of approximately 1.0 μSv a^{-1} from a new efficient plant and approximately 20 μSv a^{-1} for an older inefficient plant to a member of a critical group have been estimated [29]. The normalized collective effective doses for 1 year produced from a 1,000 MW(e) plant are as follows: 0.5 person-Sv per GW year for a typical modern plant, 6.0 person-Sv per GW year for a typical old plant, and 50 person-Sv per GW year for the typical plant operating in China, with the global average being approximately 20 person-Sv per GW year [30].

By way of comparison, radionuclides from nuclear plant operations that are released to the environment tend to be noble gases, carbon-14 and tritium. These have a different impact on uptake in the body and the effective dose received. These radionuclides are formed through fission in the uranium fuel and also by neutron activation that occurs in and around the reactor vessel. These radionuclides are released from the nuclear plant in varying levels depending on the reactor type. The most common reactor types, PWR and BWR, deliver an effective dose of 5 and 10 μSv, respectively [31]. When using global averages for all reactor types, the normalized collective effective dose is 0.43 person-Sv per GW year [31], which is slightly less than the suggested value for typical modern coal-fired power plants and significantly less than the global average of 20 person-Sv per GW year.

When comparing the radionuclide emissions and effects on the environment from nuclear power plants and coal-fired power plants, it is found that direct comparisons are difficult because of the differences in radionuclide type, plant designs, and assessment methodologies. Radionuclides emitted from coal-fired plants tend to be alkali metals and rare earths such as those from the U-238 and Th-232 series. Radionuclides released from nuclear plant operations to the environment tend to be noble gases, carbon-14 and tritium, which have a different impact regarding uptake of the radionuclides into the body and the effective dose received. McBride [27] shows the bone dose significantly greater from a coal-fired power plant compared with a nuclear plant of the same size. For the 1,000 MW(e) plants modeled the committed effective bone doses were 2.25 person-Sv for coal-fired power plants and 0.21 person-Sv for nuclear power plants. Therefore, direct comparisons can only be made when the parameters of study are carefully specified. It is certain, however, that for as long as coal-fired plants are operating there will be radionuclide releases to the environment from their operation.

Radioactive Waste The International Atomic Energy Agency (IAEA) has established a set of principles based on the basic objective that "*radioactive waste is dealt with in a manner that protects human health and the environment now and in the future without imposing undue burdens on future generations*" [8]. Radioactive waste has been disposed by many different countries in many different ways. Disposal of low-level wastes has been carried out in 20-m deep trenches and capped with a semi-impermeable cover; medium-level wastes have been disposed of in shallow burial facilities, although deep geological burial is considered more appropriate; high-level waste requires isolation from the biosphere in deep repositories [8].

There are numerous possible classifications for radioactive waste. The International Atomic Energy Agency [32] recommends a minimum waste classification system defined as follows:

1. Exempt
   Waste released from regulatory control in accordance with local exemption or clearance levels (generally, waste activity below concentrations established by government regulators).
2. Low- and intermediate-level waste (LILW)
   Generally defined as waste above exemption or clearance levels (i.e., above exempt) with low thermal powers.
3. High-level waste (HLW)
   Waste with long-lived radionuclides (>30 year half-life) above specified concentration levels and high thermal powers.

In addition to the above classifications, there are a number of descriptive waste classifications that may be utilized to further subdivide categories. Some other possible classifications [32] are listed in Table 9.

A more comprehensive approach to radioactive waste classification has been adopted [33] which incorporates defined categories as a function of waste activity and half-life. A qualitative depiction of this is presented in Fig. 9 (adapted from [33]).

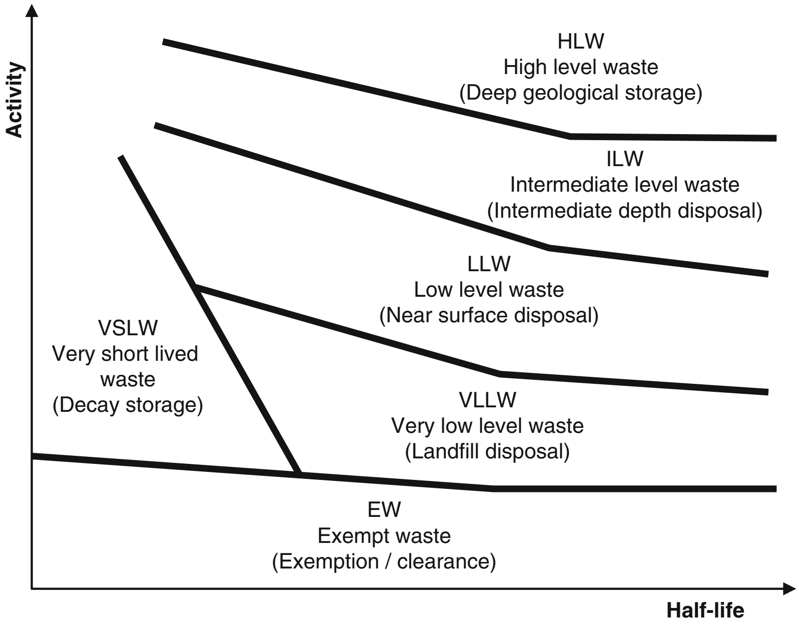
It is clear from Fig. 9 that radioactive waste may have different potential environmental impact based upon its classification and storage or disposal criteria. The lower the activity and half-life of the waste, the more immediate impact there can potentially be on the environment, due to the fact that for lower activity and half-life the material is disposed closer to human and nonhuman biota receptors. For higher activity and half-life, the disposal and storage solutions are engineered for long-term security that may last thousands of years, suggesting that any potential environmental impact will be far into the future.

There are two general types of radioactive waste disposal: (1) land disposal (or storage) and (2) ocean disposal. Land disposal may also be considered storage, since the waste will usually be in a condition whereby it could be recovered if needed. Ocean disposal of radioactive waste would generally not be retrievable due to the extreme depths involved. A third type of disposal that has been discussed in the past is nuclear waste disposal in space [34]. Investigations into this method of disposal are not being actively pursued due to high cost of disposal and potential launch risks.

All radionuclides that are found in used nuclear fuel and that are produced through nuclear power generation to be utilized for commercial, industrial, medical, or research purposes may be found in radioactive

Radiation in the Environment, Sources of. Table 9
Descriptive radioactive waste classifications

Cladding	Long lived	Primary/raw	Transuranic
Alpha bearing	Short lived	Secondary	Nuclear application
NORM	Heat generating	Mining and milling	Very low level
Mixed	Solidified	Reactor	Vitrified



Radiation in the Environment, Sources of. Figure 9
Radioactive waste classification schemes

waste. Wastes that are stored in landfills or municipal waste storage facilities are de facto introduced into the environment. However, their designation as low-level waste is strictly regulated to ensure protection of both human and nonhuman biota. High-level waste (e.g., from nuclear reactor spent fuel) has engineered barriers and its storage employs defense-in-depth strategies to prevent release to the environment.

Ocean Dumping of Radioactive Waste In the past, the disposal of radioactive waste at sea was allowed and regulated by the terms of the 1972 Convention of the Prevention of Marine Pollution by Wastes and other Matter (London Convention). In 1985, a decision was passed at the London Convention that suspended radioactive waste dumping at sea, and in 1993 the suspension was converted into an indefinite ban [8].

Historical dumping of radioactive waste into the oceans has taken place from 1946 through at least 1991 [35]. From 1946 through 1970, the USA dumped radioactive waste into the Atlantic Ocean, Pacific Ocean, and Gulf of Mexico [36]. The greatest amount of radioactive waste was, however, disposed by the former Soviet Union between the years of 1959 and 1991. Both solid and liquid radioactive wastes ranging from liquid effluent, containers, large objects, and entire vessels have been disposed into the Baltic, White, Barents, and Kara seas as well as other Arctic locations [35]. The sinking of Soviet submarines is covered in section "Nuclear-Powered Vessel Accidents". A summary of the estimated total activity introduced into the seas adjacent to the former Soviet Union is provided in Table 10 [35]. It must be noted that the activities presented represent the best estimates of activity at the time of disposal, and therefore represent a maximum bound.

There have been no measurements of dangerous levels of radioactivity in the surface seawater near any of the dumping sites [35]. However, this does not suggest that continuing monitoring and numerical prediction of release and consequences should not take place. It has been calculated that the global radiological impact of the radioactive waste dumping in the Arctic seas will be much less than that of the other forms of anthropogenic radionuclides already in existence [37]. The future impact remains to be verified through a continual environmental monitoring program near ocean radioactive waste dump sites.

Radiation in the Environment, Sources of. Table 10
Activity introduced into the maritime Arctic environment

Waste type	Activity (TBq)
Liquid radioactive waste	903
Solid radioactive waste (low and intermediate level)	590
Objects with spent nuclear fuel	85,100
Objects without spent nuclear fuel	Up to 3,700

Nuclear Weapons Testing

The world's first nuclear test detonation occurred on July 16, 1945 (Code Name: Trinity), at the Alamagordo Proving Ground Site in New Mexico, USA, and is the earliest form of anthropogenic radioactivity having potential for global environmental impact. The nuclear weapon tested was an implosion type device placed on top of a 30.5 m (100 ft) steel tower (Fig. 10). The device contained 6 kg Pu-239 and the nuclear yield was approximately 21 kT, where a kiloton (kT) is the explosive energy release expressed in equivalent tons of the chemical explosive trinitrotoluene (TNT) [38]. The explosion at 15 s post detonation and the resulting crater created at the Trinity site are seen in Fig. 11. The heat of the explosion fused elements of the device, tower, and soil into a glassy material known as Trinitite. Sixty years after the first nuclear explosion, samples of Trinitite have been analyzed [39] and determined to contain fission products (Sr-90 and Cs-137), activation products (Co-60, Ba-133, Eu-152, Eu-154, Pu-238, and Pu-241), as well as fuel remnants (Pu-239 and Pu-240). A sample of Trinitite is shown in Fig. 12. The characteristic green hue is observed, and it is noted that the side facing the heat from the nuclear device blast is smooth, whereas the side away from the blast is rough.

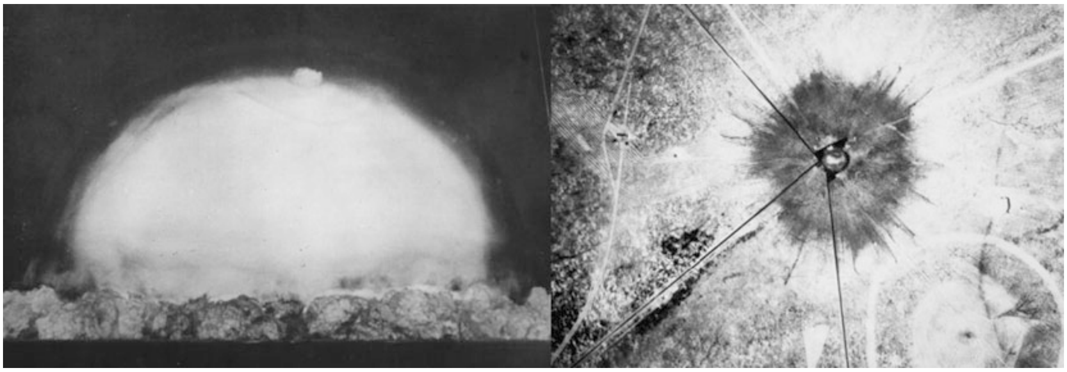
The Trinity test was the prelude for the only wartime use of nuclear weapons; when in World War II, the USA detonated two nuclear weapons against Japan: the first on August 6, 1945 (Code Name: Little Boy), and the second on August 9, 1945 (Codename: Fat Man). The Little Boy weapon was a gun-type nuclear device, producing a nuclear yield of approximately 13 kT, whereas the Fat Man weapon was an implosion device producing a nuclear yield of approximately 22 kT [40]. These first acts of exploiting the massive energy release associated with nuclear technology opened the so-called Atomic Age [41].

Radiation in the Environment, Sources of. Figure 10
Trinity tower (Photo courtesy of National Nuclear Security Administration/Nevada Site Office)

In general terms, the objectives of a nuclear test program are to ensure the operational readiness of stockpiled nuclear weapons, test new designs and to study the nature and effect of radiation fields generated by the weapons. The characteristics (such as design, weapon yield, and burst type) of nuclear weapons are important in understanding the potential distribution to and impact on the environment, and are discussed in the next section.

Characteristics of Nuclear Weapons Prior to discussing nuclear weapons testing, it is worthwhile to discuss environmental impact from the manufacture of nuclear weapons. Nuclear weapons production involves, in broad categories, the following seven steps in the fissile material production [42]:

1. Uranium mining
2. Milling



Radiation in the Environment, Sources of. Figure 11
The Trinity explosion at 15 s post detonation (*left*) and the resulting crater (*right*) (Photo courtesy of National Nuclear Security Administration/Nevada Site Office)

Radiation in the Environment, Sources of. Figure 12
Trinitite, blast side (*left*) and ground side (*right*) (Photo courtesy of E. Waller, UOIT)

3. Refining
4. Enrichment
5. Fuel and target fabrication
6. Reactor operations
7. Chemical separations

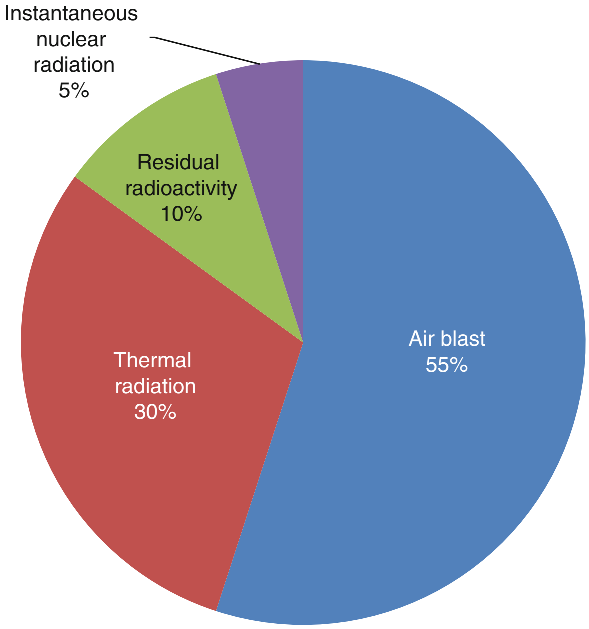
Approximately 85% of the radioactive waste generated through weapons production occurs at the chemical separations stage. For example, regarding the US Nuclear Weapons Complex, most of the environmental contamination is confined to production facilities at Hanford, Savannah River, K-25, Portsmouth, and Paducah [42] and all are under remediation.

The two principal types of nuclear weapons are (1) nuclear fission based and (2) thermonuclear fusion based. Fission weapons can be further subdivided as gun type and implosion type. The fission-based weapons were the first nuclear weapons developed and have an energy yield in the range of <1–500 kT. Subsequently, fusion-based weapons were developed (which usually involve two stages: a fission detonation rapidly followed by the fusion reaction) and capable of energy release that can exceed 50 MT.

There are numerous physical designs of nuclear weapons, which tend to depend upon the delivery system they are designated for. Different delivery systems include free-fall bomb (from aircraft), missile (as part of rocket package), torpedo (from submarine), shell (from artillery delivery), and stationary (i.e., emplacement of the weapons package statically). There are four principal categories of nuclear weapons detonation (hence testing), which include (a) exoatmospheric, (b) atmospheric, (c) underground, and (d) underwater. Each category of detonation has a different potential impact on the environment, as described below [40]:

Exoatmospheric: The primary reason to detonate in this fashion is to destroy command and communication, via disruption of satellites. First, the direct radiation can destroy sensitive electronics systems. Second, radiation generated through an exoatmospheric burst can generate a satellite phenomenon called system-generated electromagnetic pulse (SGEMP) which can destroy electronics within the satellite. Finally, it can increase the radiation density in and around the Van Allen belts, thereby increasing the probability of disabling satellite systems. The environmental impact on Earth from radiation emitted during an exoatmospheric detonation is minimal; however, the impact on the space environment can be severe.

Atmospheric: The primary reason to detonate in this fashion is to maximize damage to target(s). The spread of ground contamination is reduced with increasing height of burst from surface to near-exoatmospheric, since the interaction of the weapon radiation and material with the ground material will be reduced as the height of burst increases. The distribution of weapon energy from an atmospheric detonation at low to moderate altitudes is given in Fig. 13. Although most damage is introduced through thermal and blast effects, instantaneous neutron radiation will be responsible for activation of structural materials, and residual radiation (fallout) will be responsible for persistent environmental radiation. In addition, at very high altitude (>50 km) a very strong electromagnetic field (EMP) will be generated, which is capable of disabling electronics on ground and in the air. In addition, a number of atmospheric tests conducted by the USA (22 tests), UK (12 tests), and France (5 tests) can be classified as safety tests, or “shots” [31]. These shots were conducted to test the safety features of nuclear weapons to inadvertent detonation of the warhead (e.g., through Broken Arrow accidents). A safety shot, by design, has a desired effect of no nuclear yield, and therefore contamination should purely be a result of damage and dispersion of the weapon materials [43].



Radiation in the Environment, Sources of. Figure 13 Distribution of energy in a nuclear weapon atmospheric explosion (Adapted from [44])

Underground: This type of testing was introduced when test ban treaties were signed by the major nations testing nuclear weapons in the atmosphere. Except for weapons testing and the possibility of generating a large shock wave to destroy underground structures, there is little reason to detonate a nuclear weapon underground.

Underwater: An underwater burst can send large shock waves (both atmospheric and subsurface) to great distances, disrupting sea-going vessels of all kinds.

Nuclear radiation is emitted from both fission and fusion reactions in a thermonuclear weapon. Initial, or prompt, radiation is that generated within approximately 1 min post detonation. This includes neutrons, gamma rays, x-rays, alpha particles, and beta particles.

Only neutrons are capable of producing persistent radioactivity in the environment through neutron activation of materials. Residual radiation, primarily in the form of fission products, is manifest approximately 1 min post detonation [45]. It is this residual radiation that becomes a source of environmental radioactivity.

Global Nuclear Weapons Testing The worldwide total number of nuclear weapons tests recorded to date is 2,421, including the atomic bombs dropped on Hiroshima and Nagasaki and all safety shots [31]. A summary of the tests is provided in Table 11. Only two categories are considered: atmospheric (including all atmospheric and exoatmospheric tests), and underground (including all underground and underwater tests).

As is evidenced in Table 11, the USA and USSR conducted hundreds of nuclear weapons tests since 1945. The great quantities of measureable radioactive fallout from these tests, coupled with the unstable political situation of the 1962 Cuban Missile Crisis, lead to the development of the Limited Test Ban Treaty (LTBT) of 1963, which banned nuclear detonations in the atmosphere, exoatmosphere, and underwater. The ban did not extend to underground testing. Although the USA, UK, and USSR were signatories to this treaty, France and China were not and continued to test in the atmosphere until 1980. In 1974, the Threshold Test Ban Treaty (TTBT) came into effect and limited underground nuclear weapons tests to less than 150 kT yield. Two years later in 1976 the Peaceful Nuclear Explosions Treaty (PNET) extended the ban from weapons to include nuclear explosions for peaceful purposes. Throughout this period, a Comprehensive Test Ban Treaty (CTBT) was developed with the goal to ban all nuclear weapon testing. The treaty became available for signature in 1996, and most of the world's nations have signed (if not ratified) the treaty [46].

Radiation in the Environment, Sources of. Table 11
Summary of all nuclear weapons testing

Country	Number of tests		
	Atmospheric	Underground	Total
USA	219	908	1,127
USSR	219	750	969
France	50	160	210
UK	33	24	57
China	22	22	44
India		6	6
Pakistan		6	6
North Korea		2	2
Total	543	1,878	2,421

Of greatest importance to the injection of anthropogenic sources of radioactive material into the global environment is atmospheric nuclear testing. This is due to the fact that the fission and fusion processes generate radioactive by-products that are introduced with high energy into global atmospheric circulation patterns, and can travel aloft globally in the troposphere and stratosphere. The radionuclides distributed globally from atmospheric testing that are important from a dosimetry standpoint are presented in Table 12 [31]. The significance to environmental sources of radioactivity is clear, as many of the isotopes have been released in great abundance or have relatively long half-lives.

In the following sections, specific important nuclear weapon test sites are discussed with relevance to their environmental impact and global distribution of radionuclides.

Marshall Islands, Pacific Ocean The Pacific Proving Ground was located in The Republic of the Marshall Islands (RMI), and is composed of 1,225 islands and inlets including 29 atolls and 5 solitary, low coral islands. Centrally located in the Pacific Ocean, the Republic of Marshall Islands covers more than 5,025,000 km^2; however, only 181.3 km^2 or about 0.01% of the republic is dry land. The two primary atolls utilized for nuclear weapons testing were Bikini and Enewetak. Bikini Atoll is the largest atoll in the Republic of the Marshall Islands (RMI) and consists of 23 islands and 187 km^2 of coral reefs [47]. The atolls of the RMI were formed when fringing reefs began to establish and grow around emergent volcanoes. While the volcanic peaks slowly sank and reduced in size, the fringing reefs continued to grow and eventually coral atolls formed after the volcanoes disappeared entirely beneath the sea [47]. In the period between 1945 and 1962, a total of 5 underwater and 101 atmospheric nuclear weapon tests were conducted here.

Radiation in the Environment, Sources of. Table 12 Global release of radionuclides from atmospheric nuclear weapons testing

Radionuclide	Half-life	Global release (PBq)
H-3	12.33 a	186,000
C-14	5,730 a	213
Mn-54	312.3 d	3,980
Fe-55	2.73 a	1,530
Sr-89	50.53 d	117,000
Sr-90	28.78 a	622
Y-91	58.51 d	120,000
Zr-95	64.02 d	148,000
Ru-103	39.26 d	247,000
Ru-106	373.6 d	12,200
Sb-125	2.76 a	741
I-131	8.02 d	675,000
Ba-140	12.75 d	759,000
Ce-141	32.50 d	263,000
Ce-144	284.9 d	30,700
Cs-137	30.07 a	948
Pu-239	24,110 a	6.52
Pu-240	6,563 a	4.35
Pu-241	14.35 a	142

The Bikini and Enewetak Atolls were the first nuclear testing grounds after World War II. Starting in 1946 and ending in 1958 the USA tested 23 surface and subsurface thermonuclear devices at Bikini Atoll with an additional test conducted approximately 100 km from the atoll [48]. The total magnitude of the Bikini nuclear tests was 76.3 MT, accounting for 72% of the total yield tested in both atolls [7]. These tests caused great damage to Bikini's physical and ecological landscape creating five craters, the biggest one being the BRAVO crater which is 73 m deep [47]. The BRAVO test has been termed the "BRAVO accident" by some researchers due to the fact that the test, conducted on March 1, 1954, had a yield of 15 ± 2 MT, but prior to detonation was estimated to yield 6 ± 2 MT [49]. The creation of the BRAVO crater caused the destruction of three small islands and moved millions of tons of sand, coral, plant, and sea life from Bikini's reef [50]. Observed short-term effects in native and military populations from the BRAVO shot included nonfatal depression of blood cell production, localized skin burns from beta-emitting fallout, and internal contamination (hence dose) from fission products [49]. In addition, observance of leukemia in exposed persons has been made, as well as deleterious effects to thyroid function due to I-131 [49]. As a final consequence to the Marshallese populations living on the Bikini Atoll, major psychological trauma from repetitive relocations (due to areas being inhabitable from fallout) was observed [49].

The long-term consequences of nuclear weapons testing from chronic and large-scale anthropogenic disturbances for biodiversity are not very well understood in the Marshall Islands, to either inhabitants or the ecosystem. As a result of the nuclear weapons testing, the following principal radionuclides were released into the environment in the Marshall Islands: Cs-137, Sr-90, Pu-239,240, and Am-241 [51]. For example, soil samples from mid-Pacific global fallout are estimated to be approximately 400–800 Bq m^{-2} for Cs-137 and 0.2–0.4 Bq kg^{-1} for Pu-239,240 [48]. Distribution of radionuclides varies widely amongst the islands of the various atolls and also as a function of latitude [48]. Nuclear weapons testing also altered the natural sediment distribution by redistributing a higher amount of the fine material over the surface of the sediment, which impacts the local ecosystem [52]. The testing caused a great disturbance to the Marshall Islands environment and to local and global distribution of anthropogenic radionuclides.

Nevada Test Site, USA A number of locations within the continental USA were used for nuclear weapons testing [53], presented in Table 13. The most utilized testing site by far was the Nevada Test Site (NTS). The NTS is located approximately 105 km northwest of Las Vegas, NV. The site encompasses 3,500 km^2 and is bordered by the Nellis Air Force Range and Tonopah Test Range. The three US federally controlled areas total approximately 14,200 km^2 of unpopulated land.

After the Trinity test and the dropping of the two nuclear weapons on Hiroshima and Nagasaki in 1945, testing was moved to the Bikini and Enewetak Atolls in

Radiation in the Environment, Sources of. Table 13 Location and number of nuclear weapon tests in the continental USA

Location	Number of tests
Alamogordo, NM	1
Amchitka, AK	3
Bombing Range, NV	5
Carlsbad, NM	1
Central Nevada	1
Fallon, NV	1
Farmington, NM	1
Grand Valley, CO	1
Hattiesburg, MS	2
Nevada Test Site, NV	928
Rifle, CO	1
Total	945

the Pacific Ocean, where five nuclear weapons were tested between 1945 and 1951 [53]. Due to the large logistical effort, support, cost, and time required to perform testing at these remote locations and the requirement to accelerate nuclear weapons testing and production, a decision was made to develop a weapons testing capability within the continental USA. A number of sites were considered, and some nuclear weapons were tested at candidate sites (see Table 13). However, the primary site selected for continued testing, both atmospheric and underground, was the Nevada Test Site. The Nevada site was selected for atmospheric testing based upon its remote setting, favorable year-round weather/wind patterns, and restricted access. In addition, for underground testing, the NTS is characterized by favorable conditions for excavating deep geological trenches, slow-moving groundwater, and arid climate [54]. Atmospheric testing (199 tests) was conducted at the NTS between January 1951 and October 1958. From 1958 through 1961, nuclear testing was suspended by an international voluntary moratorium. Underground testing (approximately 809 tests) was conducted there from 1962 through 1992 (although underground testing is still allowed under test ban treaties, no nuclear weapons testing has been conducted at NTS since 1992 as the US Government adopted a second voluntary testing moratorium). Radionuclide contamination in and around the Nevada test site and neighboring states was monitored during and after each test. Numerous monitoring devices, such as ion chambers, Geiger counters, and air samplers (Fig. 14), were used to characterize the fallout. Aircraft were routinely decontaminated after flying near the plumes (Fig. 15).

The environmental radionuclide source term from the NTS has two components: atmospheric testing fallout (until October 1958) and underground testing radionuclide inventories (until October 1992). Fallout from atmospheric testing consists primarily of a large array of fission products. An estimate of the total radionuclide deposition in the continental USA from some of the more important fission products generated by NTS tests is given in Table 14 [55].

Underground testing was performed such that the majority of the radionuclides produced were contained in the deep geological formations in which the nuclear devices were tested. As such, there is no appreciable fallout. However, there is potential impact to the environment through migration of the radionuclides into the water table at the NTS. The primary radionuclide constituents involve fission products, unspent fuel actinides, tracer elements, residual tritium primarily from thermonuclear detonation, and activation products both in the weapon casing and the soil. Measurements and calculations have been performed to estimate the total radionuclide inventory below the water table at the primary NTS testing sites. An estimate of some of the principal radionuclides (decay corrected to September 23, 1992) is provided in Table 15 [49, 54].

Native American Exposure at Duckwater, Nevada A reservation was established in 1940 for the Duckwater Shoshone tribe (21 families) in Duckwater, Nevada, located approximately 160 km (100 miles) north of what was to become the Nevada Test Site (NTS). Historical wind directions from most bomb tests indicate that fallout would tend to travel predominantly north and east of the NTS, which directed fallout into populated Native American territories [56]. It has been reported that during the period of nuclear weapons testing, no health protective measures were attempted and only minimal efforts

Radiation in the Environment, Sources of. Figure 14
Two airmen check a T-33 jet airborne filter units used for collecting plume particulate (Photo courtesy of National Nuclear Security Administration Nevada Site Office)

Radiation in the Environment, Sources of. Figure 15
Decontamination of a B-50 Superfortress (Photo courtesy of National Nuclear Security Administration/Nevada Site Office)

Radiation in the Environment, Sources of. Table 14 NTS fallout radionuclides

Radionuclide	Total deposition (PBq)
^{137}Cs	2.3
^{90}Sr	1.8
^{95}Zr	220
^{103}Ru	430
^{140}Ba	1,400
^{141}Ce	500
^{144}Ce	40
^{106}Ru	24
^{89}Sr	330
^{131}I	1,500
$^{239+240}$Pu	0.13
^{241}Pu	0.54

Radiation in the Environment, Sources of. Table 15 NTS radionuclide inventory from underground testing

Radionuclide	Total deposition (PBq)
^{3}H	4,647
^{137}Cs	106
^{90}Sr	81
^{85}Kr	7
^{152}Eu	6
$^{239+240}$Pu	8
^{241}Pu	22

were made to inform people about the testing and the associated risks [57]. As a result, members of this tribe were exposed to fission products, especially volatile fission products.

A model was designed to follow a specific fission product, I-131, through a rabbit consumption scenario based upon tribal member activity. Researchers estimated that during the fall 3.2 rabbits were eaten per week, and during the rest of the year, 1.5 rabbits were eaten per week per person [57]. The pathway from the initial deposition of I-131 through to the consumption of rabbit by an adult is summarized as follows: (a) I-131 is deposited on plant surfaces from nuclear weapons testing, where it can either be reduced via weathering or by radioactive decay. (b) When a rabbit eats the vegetation, the I-131 is absorbed by the thyroid in the rabbit, with reduction occurring via radioactive decay and excretion. (c) Seasonal consumption of the rabbits by humans occurs whereby the I-131 goes to the human thyroid. (d) I-131 can also be transferred by breast milk to infant. In addition, the mobility of the Duckwater Shoshone tribe members was taken into account, insofar as they tended to hunt over a wide area. Twenty-five test events were reported to have deposited fallout in Duckwater or within 50 miles of Duckwater [57]. Dose estimates and thyroid cancer excess relative risk (ERR) estimates were derived based upon the largest events. Elevated ERR estimated were generated for the vulnerable population (<15 years of age when exposed), although there is no epidemiological data to support excess observed thyroid cancers to date. Of interest to the subject of sources of radioactivity in the environment, this study depicts an environment exposed to multiple fission products, extending numerous decades of weapons testing, which has elements of vulnerable populations (children), and has relatively stationary ecosystem members (rabbits), as well as highly mobile target populations (adults), which all point to the complexity of establishing environmental pathways for identified groups. Significant whole body exposures (upward of 10 mSv) have been estimated to residents of Duckwater and other Native American communities in the environs surrounding the NTS. As well, estimates of up to 200 mSv thyroid dose (from I-131) to these populations have been made [56].

Maralinga Test Site, Australia The UK conducted nuclear weapons testing between 1952 and 1957 on the Australian continent. The tests were conducted at Monte Bello Island, Western Australia (3 tests), Emu Field, Southern Australia (2 tests), and Maralinga Range, South Australia (7 tests). The nuclear weapons tests ranged in nuclear yield from approximately 1 to 60 kT. In addition to the above nuclear weapons tests, the UK conducted a number of "minor trials" designated primarily as safety shots, producing little to no nuclear yield. The safety shots were designed to explore the potential dispersion of material from a nuclear weapon under conditions of burning and explosive dispersal.

The most used test site was the Maralinga range, which hosted seven nuclear weapon detonations and hundreds of safety shots. The radionuclides released from the weapons tests were standard fission and activation products, many of which rapidly decayed away. However, the safety shots generated large areas of plutonium and uranium contamination in Maralinga, and specifically the area of Taranaki became the most contaminated region in Maralinga [58]. The principal limiting radionuclide for environmental exposure is Pu-239 and studies have been performed to characterize the plutonium contamination [59]. Soil concentrations around Maralinga are estimated on average at a level of 10 Bq g^{-1} [60] and annual effective doses to members of critical groups (typically aboriginal peoples residing in the area) have been conservatively estimated to approach 0.5 Sv [58]. Rehabilitation of some of the areas at Maralinga (including Taranaki) has been performed by collecting contaminated soils, placing in trenches, and covering with 5 m of clean soil. The Maralinga site rehabilitation effort is ongoing [61].

Mururoa, French Polynesia Early (1960–1966) French nuclear weapon testing (17 tests) was conducted in French Algeria [62]. As a result of Algeria gaining independence in 1962, the French moved testing to the Mururoa and Fangatuafa Atolls in French Polynesia. A total of 178 nuclear weapon tests were conducted in the Atolls along with 15 safety shots at Mururoa [63], as presented in Table 16. It may be seen that Mururoa was the most utilized test location.

The Atolls are largely uninhabited and as a result there was little potential impact to human exposure. Studies have been performed to quantify the levels of certain fission products such as Cs-137 [64] and actinides such as Pu-239 and Pu-240 [65]. In general, long-lived fission products and actinides have been observed in the lagoon sediments. The largest potential contributor to environmental exposure is plutonium, which is decreasing over time due to removal from the sediments and dilution in a large Pacific Ocean volume [66]. Tritium has also been measured in the lagoons; however, it is estimated that any dose contribution from the tritium is negligible [66]. The estimated inventories of some important radionuclides from the lagoons of the Atolls are presented in Table 17 [67].

Radiation in the Environment, Sources of. Table 16
French nuclear weapon tests

Location	Atmospheric	Underground	Total
Mururoa	37	127	178
Fangataufa	4	10	
Safety shot	5	10	15

Semipalatinsk Test Site, Kazakhstan The former Soviet Union (USSR) conducted nuclear tests from 1949 to 1990 primarily at the Semipalatinsk test site (456 tests) in Kazakhstan and the Novaya Zemyla test site (130 tests) in the Arctic Archipelago [68]. In addition, three other locations hosted nuclear weapons tests, specifically (1) missile test range, Astrakhan region; (2) firing grounds, Soviet Ministry of Defense, Totsk; and (3) near Aralsk. The former Soviet Union conducted peaceful nuclear explosions (119 tests) in a number of locations around the former Soviet Union. All nuclear weapon and peaceful use tests are outlined in Table 18 [68].

It is clear that the Semipalatinsk test site was the most utilized of all nuclear test sites. A total of 111 atmospheric detonations (86 air and 25 surface) were conducted between 1949 and 1963 at Semipalatinsk, causing the majority of contamination to the environment and most exposure to the public. Four of the tests, occurring in 1949, 1951, 1953, and 1956, are estimated to have contributed to more than 95% of

Radiation in the Environment, Sources of. Table 17
Radionuclide inventories from Mururoa and Fangataufa Atolls

Radionuclide	Total activity (TBq)
$^{238+239+240}Pu$	30
^{241}Am	<1
^{137}Cs	<1
^{152}Eu	<1
^{60}Co	<1

the dose to the population surrounding the Semipalatinsk test site [70].

Considerable levels of contamination can be found in and around the Semipalatinsk test site, as demonstrated in Table 19 [71, 72]. “Ground Zero” refers to the area immediately under the first thermonuclear test in 1953, and Chagan (sometimes referred to as Bolapan) refers to a crater formed by a 1965 underground nuclear explosion. The total inventory of radionuclides important as environmental contaminants near the Semipalatinsk test site is provided in Table 20 [68].

Radiation in the Environment, Sources of. Table 18 Location and number of nuclear weapon tests in the former USSR

Location	Number of tests
Semipalatinsk	456[a]
Novaya Zemlya	130
Russian federation (European)	59
Russian federation (Asian)	32
Ukraine	2
Kazakhstan	33
Uzbekistan	2
Turkmenia	1
Total	715

[a] This number is sometimes quoted as 498, which takes into account unsuccessful nuclear explosions that contributed to spread of contamination [69].

Numerous studies have been conducted on the population and ecology of the former Soviet Union, especially near the Semipalatinsk test site. Radionuclides have been detected in the terrestrial environment, members of the food chain, and efforts have been made to retrospectively calculate the dose to surrounding populations (see, e.g., [73, 74]). The investigations will continue for many decades.

Nuclear Material Accidents

There are six main categories of nuclear accidents considered herein with respect to release of radioactive material to the environment:

1. Nuclear reactor accidents
2. Fuel processing accidents
3. Nuclear weapon accidents
4. Nuclear-powered vessel accidents
5. Depleted uranium munitions
6. Space nuclear power source accidents

Accidents and incidents involving medical, industrial, and lost/orphan sources are discussed in their respective sections. Although not accidents per se, depleted uranium munitions use is included herein since the controversy surrounding depleted uranium use in the battlefield was not anticipated a priori.

It is worthy to note that, due to the secrecy surrounding nuclear reactor and nuclear weapon capability in many countries, there may be many more undocumented minor releases to the environment or lost nuclear weapons. It is expected that most major releases to the environment will be detected by

Radiation in the Environment, Sources of. Table 19 Radionuclide contamination (Bq kg^{-1}) in and around Semipalatinsk

Radionuclide	On-site [71]		Nearby village [72]
	Ground zero	Chagan	Dolon
^{137}Cs	108–83,300	1,500–22,600	2.8–71.4
^{60}Co	<2–5,410	356–21,840	–
^{152}Eu	17–96,100	321–17,200	–
^{154}Eu	<6–2,910	585–11,430	–
$^{239+240}$Pu	488–27,900	192 – 8,850	1.7–94.9
^{241}Am	52–520	24–1,056	–

appropriate metrology, although the potential impact of historical releases that have not been documented may never be known, or their impact has not yet been realized.

Nuclear Reactor Accidents Nuclear reactor accidents are amongst the least common of all industrial accidents. There have only been a few accidents that have resulted in radioactive releases to the environment and/or death, compared with environmental impact and accidents from other industries such as chemical processing, mining, or manufacturing.

A list of nuclear reactor accidents contributing to environmental source terms is provided in Table 21. The table only describes accidents where there were measured releases of radioactive material to the environment, and not core releases that were contained within the reactor environs.

Three Mile Island Unit 2 The interest in the TMI-2 accident is not as much in its environmental releases (which were minimal) but in the damage it did to the nuclear industry in the USA and abroad. The accident has been well analyzed [75], and the details will not be discussed herein.

On March 28, 1979, a combination of equipment and operator errors culminated in the core of the reactor partially (approximately 50%) melting. This led to stack venting of built-up noble gases and volatile iodine [25]. The activities of principal radionuclides released as a result of the TMI-2 accident are shown in Table 22 (adapted from [2]). There is no environmental impact from the releases since all principal radionuclides have short half-life. In addition, there was no significant radiation exposure to any workers or members of the public.

Chernobyl NPP Unit 4 Special discussion must be given to the Chernobyl (also known as Chornobyl) Nuclear Power Plant Unit 4 accident as it is responsible for the single greatest release of anthropogenic radioactive material into the environment from a nuclear reactor accident. A great number of essays, treatises, and books have been written about the technical, political, sociological, and health impacts of the Chernobyl accident, and as such this section makes no attempt to cover these issues in any detail. A brief summary of the accident follows, with an overview of the sources of radiation introduced into the environment from this accident.

The Chernobyl Nuclear Power Plant (NPP) is located in Northern Ukraine, in relatively close

Radiation in the Environment, Sources of. Table 20 Total radionuclide inventory of important radionuclides near Semipalatinsk

Radionuclide	Total activity (TBq)
$^{238+239+240}$Pu	0.15
^{90}Sr	1
^{137}Cs	1.5

Radiation in the Environment, Sources of. Table 21 Nuclear reactor accidents with environmental releases

Date	Reactor	Release to environment
December 12, 1952	NRX, Chalk River, Ontario, Canada	Fission products, principally Pu, Xe, Te, I, Cs and Ru
October 10, 1957	Windscale Pile 1, Cumbria, UK	Fission products, principally I-131, Cs-137, Sr-89,90 and Po-210
July 26, 1959	Sodium Reactor Experiment, Santa Barbara, California, USA	Radioactive gases released from containment
January 3, 1961	SL-1, near Idaho Falls, Idaho, USA	I-131
January 21, 1969	"Experimental nuclear reactor," Lucens Vaud, Switzerland	Reactor was housed in a natural cavern under a hill. No radioactivity was released from cavern
March 28, 1979	TMI-2, Middletown, Pennsylvania, USA	Noble gas fission products and I
April 26, 1986	Chernobyl Unit-4, Chernobyl, USSR (now Ukraine)	Fission and activation products, discussed below

Radiation in the Environment, Sources of. Table 22
Estimated activity of radionuclides released from 1979 TMI-2 accident

Nuclide	Activity (PBq)
^{88}Kr	1.4
^{133}Xe	55.5
^{133m}Xe	8.5
^{135}Xe	11.1
^{135m}Xe	0.9
^{131}I	0.00056

proximity to both Russia and Belarus. The nuclear plant site is situated near the Pripyat River, which feeds into the larger Dnieper River that flows south to the capital city of Kiev. The site hosted six nuclear plants: units 1–4 were operational, whereas units 5 and 6 were under construction at the time of the accident. The reactors were of the RMBK type, which are water-cooled, graphite-moderated reactors. Each reactor was capable of producing approximately 3,200 MW_t, or 1,000 MW_e energy. The reactor was designed with no containment, which suggests that under a severe accident scenario, there would be no barrier between the radioisotopes produced in the fuel and the environment. Unfortunately, in the early hours of April 26, 1986, the Chernobyl NPP Unit 4 had a "beyond design basis" accident, which included a power excursion to at least 300,000 MW_t that produced two major steam explosions within the reactor, blowing the 1,000 t top plate off, destroying all 2,000 fuel channels and releasing massive quantities of radioactive material into the environment.

Activity of the principal radionuclides released in the accident may be found in Table 23. The values presented in the table are averages of a number of estimates made in 1996 [31].

Although only about 3.5% of the core was ejected from the reactor during the accident, extremely large quantities of fission products were released into the atmosphere with much of it being distributed globally. In addition, high percentages of core inventory of certain volatile radionuclides were released, notably approximately 30% of Cs-137 and 50% of I-131 core inventories [31]. The principal dose-producing radionuclides emitted from the Chernobyl accident were Xe-133, I-131, Cs-137, Sr-90, and Pu-239 [25]. Of those, Xe-133 and I-131 were dispersed quickly and have short half-lives, thereby mitigating their concern as long-term environmental contaminants. The Pu-239 and Sr-90 were found to predominantly deposit in the near (<30 km) vicinity of the release, and therefore may be considered a local environmental concern (although most of the contaminated soil was stripped and stored at the Chernobyl NPP4 site). Cesium-137 is significant because of its substantial contribution to lifetime effective dose, its relatively long radiological half-life (30.2 years), and its ease of measurement in

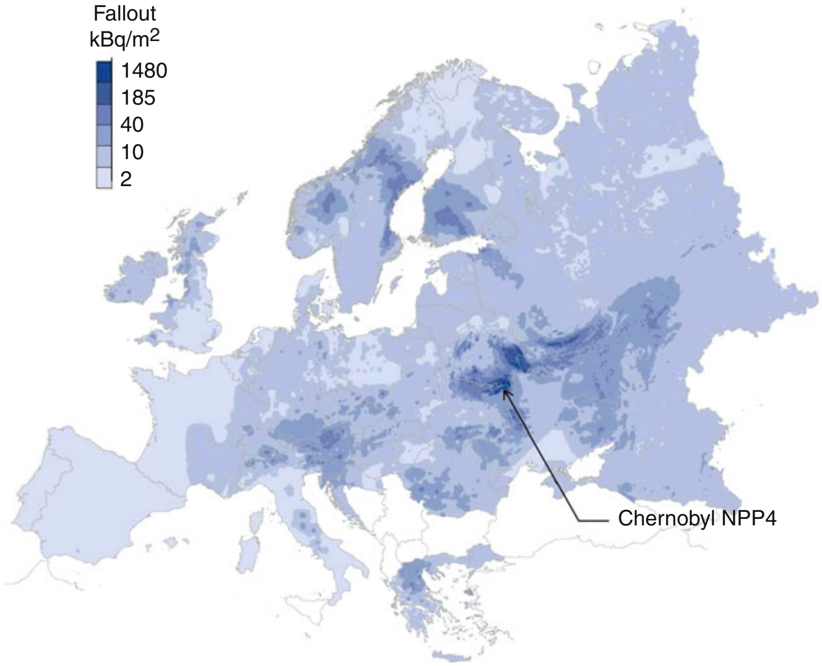
Radiation in the Environment, Sources of. Table 23
Estimated activity of radionuclides released from 1986 Chernobyl accident

Radionuclide	Activity (PBq)
^{85}Kr	33
^{89}Sr	92
^{90}Sr	9
^{99}Mo	189
^{95}Zr	169
^{103}Ru	153
^{106}Ru	43
^{131}I	1,753
^{133}I	2,500
^{132}Te	1,075
^{133}Xe	6,500
^{134}Cs	51
^{137}Cs	85
^{140}Ba	190
^{141}Ce	172
^{144}Ce	115
^{239}Np	1,323
^{238}Pu	0.033
^{239}Pu	0.031
^{240}Pu	0.046
^{241}Pu	6
^{242}Cm	1

environmental samples (through gamma spectroscopy). Therefore, Cs-137 has been historically chosen as the key indicator radionuclide for environmental contamination [31].

Ground deposition of radioactive material was found in virtually every country in the northern hemisphere. No contamination due to the accident was detected in the southern hemisphere, although interhemispheric transport of contamination occurred through a small extent from human activities, such as shipping materials and food [31]. The predominant radionuclide found globally in the environment from the accident is Cs-137, due to its volatile nature and long half-life. Iodine-131 was important in the early stages of the accident, due to its volatile nature and its impact on the human thyroid. However, with a short half-life of 8 days, it is not a persistent environmental radionuclide. Contributions from Chernobyl to local distributions of Cs-137 have been detected and confirmed in the environment using ratios of Cs-134 to Cs-137 (e.g., [76]). To illustrate the widespread dispersion of radioactivity into the environment, the distribution of Cs-137 around Europe in terms of surface contamination is depicted in Fig. 16. Areas of elevated contamination around the accident site are observed, as well as other isolated areas generated by atmospheric deposition around the time of the accident. Specifically, Finland, Norway, Sweden, Austria, and Bulgaria all recorded elevated levels of Cs-137 [31]. Prior to the Chernobyl accident, all environmental Cs-137 was due to nuclear weapons fallout.

In the area around Chernobyl NPP4, known as the exclusion zone, environmental sampling is conducted to help assess the long-term impact of the accident. Soil and vegetation sampling (Fig. 17) routinely takes place to monitor radionuclide (specifically Cs-137) movement and soil-vegetation transfer factors. Measurable quantities of long-lived radionuclides may be found in local biota. Figure 18 shows a beta particle detector reading above background levels on a moose bone. The beta radiation is likely from Sr-90, which has a half-life of 30 years, is a strong beta emitter and was released in abundance from the accident. Strontium acts similar to calcium when taken into the human body, and tends to accumulate where calcium accumulates, hence the elevated bone reading.



Radiation in the Environment, Sources of. Figure 16
Distribution of Cs-137 around Europe in May 2006 (Figure modified from [77] based on original from EU/IGCE)

Radiation in the Environment, Sources of. Figure 17
Environmental sampling in the Chernobyl NPP4 exclusion zone (Photo courtesy of E. Waller, UOIT)

Radiation in the Environment, Sources of. Figure 18
Beta radiation readings on bone fragment (Photo courtesy of E. Waller, UOIT)

The entrance to the exclusion zone is strictly controlled by monitoring entrance and exit points, which are under control of shelter maintenance and emergency services (Fig. 19 depicts the signage circa 1999). This mitigates problems surrounding movement of contamination to areas outside of the exclusion zone. Regarding the radiological state of the reactor, in mid-May 1986 the decision was made to cover the destroyed Unit 4 with a barrier to prevent further environmental contamination and improve the radiological situation of the Chernobyl NPP4 site [78]. In just over 200 days, by the end of November 1986, the massive concrete and

Radiation in the Environment, Sources of. Figure 19
Sign entering Chernobyl (Photo courtesy of E. Waller, UOIT)

steel structure, known as the "shelter" or "sarcophagus" was constructed. Figure 20 shows the shelter (darker gray) over the original structure (lighter gray).

The shelter was never intended to be a permanent structure, and as such has been in a constant state of deterioration since initial fabrication. As a result, in the mid-1990s plans were made to construct a new permanent structure. The goals of this structure, termed the "New Safe Confinement," were to (a) transform Unit 4 into an environmentally safe system, (b) reduce weathering and corrosion of the existing Unit 4 and shelter structure, (c) mitigate the consequences of a Unit 4 structural collapse, and (d) facilitate safe deconstruction of the unstable Unit 4 structures. The New Safe Confinement is designed to be emplaced over the existing Unit 4 and shelter. It is currently scheduled for completion in 2012.

Fukushima Daiichi Units 1–6 On 11 March 2011, an earthquake and subsequent tsunami off the east coast of Japan caused destabilization of a number of nuclear reactors north of Tokyo, resulting in the release of radioactive material to the environment. This accident will be discussed in the section "Future Directions" – Risk Assessment.

Fuel Processing Accidents There have been a small number of significant environmental releases of radioactivity associated with nuclear fuel processing/reprocessing facilities. In November 1959, there was a chemical explosion at the Oak Ridge National Laboratory radiochemical processing pilot plant, which contaminated the pilot plant building and nearby streets with plutonium [25]. Although the plutonium escaped the confines of the plant, the area was cleaned and no persistent environmental contamination ensued. On September 30, 1999, a prompt criticality occurred at the JCO fuel processing facility in Tokaimura, Japan, resulting in the deaths of two

Radiation in the Environment, Sources of. Figure 20
Chernobyl NPP4 shelter (circa 1999) (Photo courtesy of E. Waller, UOIT)

workers and small releases of radioactivity (principally I-131 and noble gases) to the nearby environment. Due to the short half-lives of the radionuclides released, there is no persistent impact on the environment. A number of other process criticality accidents have occurred over the past 60 years [79] that have produced minor venting of radionuclides to the environment. Two separate facilities in the former Soviet Union/ Russian Federation are worthy of special mention here, due to the significant environmental releases associated with both: Siberian Chemical Enterprises in Tomsk, and the MAYAK Production Center in the Ural Mountains.

Tomsk-7 An accident occurred at a Siberian Chemical Enterprises nuclear fuel reprocessing plant, known as Tomsk-7, located near Tomsk in the Russian Federation on April 6, 1993. An overpressure occurred in a tank containing uranium nitrate and a number of radionuclides. The most likely cause of the accident was improper mixing of various solutions that were present in the tank, causing stratified layers of organic material, radioisotopes and uranium nitrate [80]. Nitric acid introduced into this non-mixed solution caused oxidation and nitration of the organic layer (so-called red oil), ultimately resulting in an exothermic reaction [81]. This exothermic reaction increased the tank pressure to 17 atm and resulted in the tank exploding. As a result, approximately 10% of the tank contents were released to the environment, contaminating approximately 135 km^2. The estimated activity and composition of the radionuclides released is given in Table 24.

There was environmental impact both in the environs around the plant, and extending outward to the nearby village of Georgievka and to a lesser extent Naumovka [80]. Relatively low-level contamination was measured in both air and to a lesser extent surface water samples taken around the time of the accident near the accident site. It is estimated that up to 30% of the fallout activity deposited on forest canopies, in the same ratio as presented in Table 24. Many buildings, roads, and infrastructure around the Tomsk-7 site were decontaminated, and waste was buried in a disposal site at Tomsk-7. The relatively short half-lives of the principal radionuclides (except for Pu-239), ensured that the radioactivity associated with the waste dropped off fairly rapidly. Radioactive contamination

Radiation in the Environment, Sources of. Table 24 Estimated activity of principal radionuclides released from the 1993 Tomsk-7 accident

Radioisotope	% Composition in fallout	Estimated activity released (TBq)
^{95}Zr	20.4	6.5
^{95}Nb	44.0	14.3
^{103}Ru	1.4	0.36
^{106}Ru	31.4	9.5
^{125}Sb	0.4	0.1
^{141}Ce	1.5	0.37
^{144}Ce	0.9	0.24
^{239}Pu	0.01	6.3e-3

Source: Adapted from [80].

(of the composition presented in Table 24) was measured around the neighboring village of Georgievka, and this area was also decontaminated. The long-term impact of the accident is limited to the environmental deposition of the longest-lived radionuclide, Pu-239, which was a small percentage of the total fallout. The major constituents of the fallout have relatively short half-lives and decayed fairly rapidly.

Mayak The Mayak Production Association, located in the Chelyabinsk region of the southern Ural mountains, was opened in June 1948 as a critical component of the nuclear weapons complex in the former USSR. The primary mission of the facility was to produce plutonium (Pu-239) for nuclear weapons. The Mayak facility encompassed 90 km^2 surrounded by a buffer zone of 250 km^2 [82]. The facility consisted of uranium-graphite reactors (seven reactors over the years spanning from 1948; five reactors decommissioned), radiochemical plants, and waste storage. Since 1962, Mayak has reprocessed nuclear fuel and has been a significant supplier of radioisotopes. A number of accidents and incidents have occurred over the years at this facility that have resulted in significant releases to the environment and significant dose to both Mayak workers and the general population around the facility.

In the time frame 1951–1967, gaseous aerosol releases in the form of noble gases, I-131, and other volatile materials were emitted from the reactor and radiochemical plant side. Also within this time frame, both planned and unplanned liquid waste releases were introduced into the Techa River from facility operation. A significant accident occurred on September 29, 1957, at the Kyshtym high-level waste storage tank. An equivalent 100 t TNT chemical explosion resulted in significant spread of contamination over 20,000 km^2 in a narrow band extending from Kyshtym to the north eastern village of Kamyshlov [59]. In addition to the above, a number of "wild overflows" occurred in the liquid waste storage tanks resulting in significant environmental contamination of the Techa river. In 1951 an open liquid waste reservoir known as Karachay Lake began operation. In 1967, during a hot spring following a dry winter, the lake partially dried up and large quantities of radioactive sediment underwent wind-driven resuspension and were transported away from the lake. It is impossible to determine with certainty all of the releases that have occurred with respect to operation of the Mayak facility, due to the secrecy of the operation, incomplete record keeping, extended timeline over which the releases occurred, and absolute magnitude of the releases. Some of the more significant environmental releases, and approximate radioisotope activities, are presented in Table 25.

International Nuclear Event Scale To aid in determining the severity of nuclear and radiological incidents, the International Atomic Energy Agency (IAEA) devised the International Nuclear and Radiological Event Scale (INES). The purpose of the INES is to facilitate communication between the technical community, the media and the public on the significance of nuclear or radiological events [83]. The rating system takes into account the impact on workers, the public and the environment. The position on the scale is determined from the highest of three numerical scores from examining on-site effects, off-site effects, and degradation of the defense-in-depth principle. The INES scale is depicted in Fig. 21. Some of the higher INES-level events are the Chernobyl NPP4 accident (INES Level 7); the Mayak Kyshtym explosion (INES Level 6); the Windscale fire, TMI-2 accident, and Goiania accident (INES Level 5); and the SL-1 accident and Tokaimura accident (INES Level 4).

Radiation in the Environment, Sources of. Table 25
Estimated releases to the environment from the Mayak Production Association facility

Release	Date(s)	Principal radioisotope(s)	Activity released (PBq)
Gaseous aerosol releases	1951–1967	^{131}I	20.0
Liquid waste releases including wild overflows	1949–1956	Rare earth elements	27.0
		^{103,106}Ru	26.0
		^{95}Zr/Nb	14.0
		^{137}Cs	12.0
		^{89,90}Sr	20.0
Kyshtym accident	September 29, 1957	Rare earth elements	70.0
		^{90}Sr	3.75
Karachay Lake Resuspension	April–May 1967	^{137}Cs	0.011
		^{90}Sr	0.004

Source: Adapted from [8, 82].

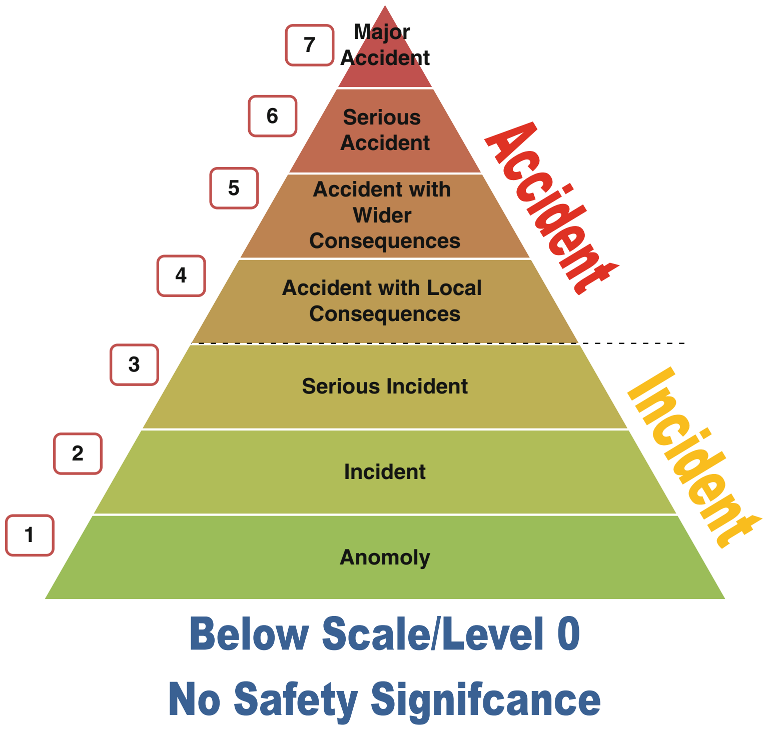
Nuclear Weapon Accidents There have been relatively few nuclear weapon accidents or incidents since the development of nuclear weapons in 1945. The major threat from an accident involving a nuclear weapon is not a nuclear detonation, but rather a spread of contamination through the environment via explosion and/or fire associated with an accident. The reason for this is simply that nuclear weapons require a number of systematic initiation steps including physical insertion of materials and setting of interlocks prior to their use as weapons. Also, typically, accidents occur when weapons are being ferried in a not fully functional state or are in an otherwise relatively inert state. That being said, nuclear weapons contain significant quantities of conventional explosives which can detonate in an accident, spreading the nuclear material (e.g., U-235, Pu-239, H-3, and depleted uranium) into the environment. For the purpose of this section, discussion will include accidents involving nuclear weapons (bombs, missiles, etc.). Nuclear-powered sea-going vessels that also often serve as nuclear weapon platforms will be discussed in section on "Nuclear-Powered Vessel Accidents".

Unplanned events involving nuclear weapons can be categorized as either accidents or incidents. A nuclear weapon accident can be broadly categorized as unauthorized launching, unexpected detonation or burning, release of contamination from an event, or theft/loss. A nuclear weapon incident can be broadly categorized as an event leading to increased possibility of explosion or contamination, errors/malfunction in the weapon system (not leading to an accident), or any force majeure resulting in damage of a weapon, facility, or component. In terms of severity, a nuclear accident is more severe than a nuclear incident. Similar to the INES accident severity scale (Fig. 21), the US Navy devised further subdivisions to clarify the magnitude of the event. These subdivisions are as follows:

1. Nucflash: Any accidental or unauthorized incident involving a possible detonation of a nuclear weapon by US Forces that could create the risk of nuclear war.
2. Broken Arrow: The accidental or unauthorized detonation, or possible detonation of a nuclear weapon (other than war time) including the non-nuclear detonation or burning of a nuclear weapon; radioactive contamination; and seizure, theft, or loss of a nuclear weapon or component (including jettisoning).
3. Bent Spear: Any nuclear-weapon-significant incidents other than accidents or war detonations.
4. Dull Sword: Any nuclear weapon incident other than significant incidents.
5. Faded Giant: Any nuclear reactor or radiological accidents involving equipment used in connection with naval nuclear reactors or other naval nuclear energy devices while such equipment is under the custody of the Navy.

In this section, only events related to introduction or release of radioactivity to the environment are considered, and are therefore classified as Broken Arrows.

Accidents Involving Nuclear Weapons The majority of accidents involving nuclear weapons have occurred during military aircraft operations during the Cold War between the USA and former Soviet Union (1947–1991). Due to extreme secrecy of the former



Radiation in the Environment, Sources of. Figure 21
INES Scale (Adapted from [83])

Soviet nuclear weapons program, there is very little known in the open literature regarding Soviet nuclear weapon accidents, and therefore are not discussed herein. The most prominent US nuclear weapon accidents (Broken Arrows) involving release of radionuclides to the environment are listed below:

1. B-36 bomber ditching with nuclear weapon off the coast of British Columbia (February 13, 1950). A B-36 bomber carrying a nuclear weapon with a “dummy” warhead (uranium instead of plutonium) ditched off the coast of British Columbia. The nuclear weapon was jettisoned and the weapon’s conventional explosives detonated on impact with the ocean.
2. B-50 bomber jettison and detonation of nuclear weapon near Rivière-du-Loup, Québec (November 10, 1950). A B-50 bomber upon having engine problems over the St. Lawrence River in Canada jettisoned its nuclear weapon which was set to self-destruct at 760 m (2,500 ft). There was no plutonium in the weapon; however, approximately 45 kg (100 lb) of depleted uranium was dispersed around Rivière-du-Loup.
3. B-47 bomber loss over the Mediterranean Sea (March 10, 1956). Two capsules of nuclear weapons material were on board the aircraft when it disappeared. The aircraft was never found, and was presumed to have crashed into the sea. The states of the two nuclear weapons capsules are unknown.
4. B-36 bomber inadvertent jettison of weapon at Kirtland Air Force Base, New Mexico (May 27, 1957). A (non-armed) nuclear weapon was inadvertently released at low altitude, and upon striking, the ground was destroyed through detonation of its convention explosives. Low levels of radioactivity were detected in the impact crater.
5. C-124 Globemaster II cargo aircraft jettison of weapons off the coast of New Jersey (July 28, 1957). Three nuclear weapons and one fissile core were being transported when the aircraft

lost power and the crew jettisoned two of the nuclear weapons. The states of the two nuclear weapons are unknown.

6. B-47 bomber take-off roll accident and fire in Morocco (January 13, 1958). A B-47 performing a simulated take-off while carrying a nuclear weapon experienced a ruptured fuel tank and subsequent fire. The nuclear weapon high explosives did not detonate, although low-level alpha contamination was detected in the vicinity of the accident.
7. B-47 bomber jettison of weapon off the Atlantic Coast near Savannah, Georgia (February 5, 1958). A mid-air collision damaged the bomber, forcing it to jettison the single nuclear weapon it was carrying over the Atlantic Ocean. The state of the nuclear weapon is unknown.
8. C-124 Globemaster II cargo aircraft destruction of weapon at Barksdale Air Force Base, Louisiana (July 6, 1959). The transport aircraft crashed on take-off carrying one nuclear weapon. The fire destroyed the aircraft and weapon, and a small amount of radioactive contamination was discovered in the immediate vicinity of the crash.
9. BOMARC nuclear missile destroyed by explosion at McGuire Air Force Base, New Jersey (June 7, 1960). A fire at a BOMARC air defense missile facility lead to the destruction of the nuclear payload, leading to release of plutonium to the atmosphere and local contamination of the site.
10. B-52 bomber mid-air explosion loss of weapons near Goldsboro, North Carolina (January 24, 1961). A B-52 carrying two nuclear weapons suffered a mid-air explosion, causing its two nuclear weapons to come free during the aircraft breakup. One weapon was recovered relatively intact. The other weapon broke up upon contact with the ground. Most of the thermonuclear stage and enriched uranium was never located, and the area of impact was purchased by the US Air Force and secured. The area is routinely monitored for evidence of environmental contamination.
11. Thor rocket with nuclear payload destruction near Johnson Atoll, Pacific Ocean (June 4, 1962). A Thor rocket carrying a nuclear device for a high-altitude nuclear test was destroyed after control of the rocket was lost. It is believed that the nuclear device vaporized through the destruction. The nuclear material would have been dispersed through the atmosphere and mostly settled in the Pacific Ocean.
12. Thor rocket with nuclear payload destruction near Johnson Atoll, Pacific Ocean (June 20, 1962). A Thor rocket carrying a nuclear device for a high altitude nuclear test was destroyed after the booster rocket shutdown prematurely. The nuclear device is believed to have fallen into the Pacific Ocean. The state of the device is unknown.
13. B-58 bomber accident and destructive ground fire at Grissom Air Force Base, Indiana (December 8, 1964). A B-58 bomber slid off a taxiway and caught fire with five nuclear weapons on board, which were damaged as a result. Contamination was detected and confined to the immediate vicinity of the accident. A B-58 Hustler, exhibiting the large nuclear weapons payload capability, is shown in Fig. 22. A large nuclear device (multi-color weapon near bottom center of Fig. 22) is wrapped within the fuel pod (termed the “two component pod”), and the four smaller weapons are attached directly to hard-points under the wings. Operationally, after the final refueling before target, the aircraft would burn the fuel from the pod tank and jettison the pod from the aircraft, leaving the large nuclear weapon for delivery. Therefore, an aircraft fully fueled prior to take-off in this configuration would be of great concern in the event of fire on the ground.
14. A-4E attack aircraft lost from carrier with nuclear weapon near the Ryukyu Islands, Japan (December 5, 1965). An A-4E armed with a nuclear weapon rolled off the deck of the USS Ticonderoga and sank to a depth of approximately 4,875 m (16,000 ft). The state of the weapon is unknown.
15. B-52 bomber with nuclear weapons mid-air collision near Palomares, Spain (January 17, 1966). A B-52 collided while refueling with a KC-135 tanker, destroying both aircraft. The conventional explosives on two of the four nuclear weapons detonated upon impact with the ground spreading low levels of radioactive contamination over the countryside around Palomares. See detailed description below.

Radiation in the Environment, Sources of. Figure 22
B-58 Hustler nuclear capable strategic bomber (Photo courtesy of Lt. Col. BJ Brown, US Air Force (Ret))

16. B-52 bomber destruction while carrying four nuclear weapons at Thule Air Force Base, Greenland (January 21, 1968). A fire in the navigator's compartment caused a B-52 to attempt an emergency landing. The aircraft crashed in the attempt and the fire caused at least one of the nuclear weapon's conventional explosives to detonate. The explosion spread contamination (primarily plutonium) around the accident site. Although there are conflicting reports, it is anticipated that four of the weapons were destroyed. There is conjecture that one of the weapons actually fell through the ice to the ocean floor, but this is unsubstantiated.

It is clear that there are two generic categories of Broken Arrow accidents: (1) lost and (2) destroyed nuclear weapons. The environmental impact of a destroyed nuclear weapon is immediate, whereas the impact of a lost weapon in unknown, with the impact possibly occurring at some time in the future (note that lost weapons may be destroyed, but with no way to confirm). In all destroyed weapon accidents, there were decontamination efforts to mitigate or eliminate potential effects. A summary of the US nuclear weapon accidents resulting in either lost radioactive material or destruction of weapon is presented in Table 26.

Accident at Palomares, Spain The Palomares Broken Arrow incident is of special interest since it was responsible for widespread radionuclide contamination of the environment. On January 17, 1966, a B-52 bomber of the US Strategic Air Command collided with a KC-135 tanker while trying to perform an airborne refueling operation at 31,000 ft above Palomares, Spain. The collision destroyed both aircraft, and the four Mk-28 thermonuclear weapons were expelled from the aircraft during the B-52 breakup. Three of the weapons were recovered within 24 h on land (of which two had undergone detonation of their high explosive charges, dispersing radioactive material into the environment), while the fourth weapon landed at sea and was the focus of an extensive deep sea search and recovery operation [84]. The fourth weapon was recovered relatively intact with no release of radioactivity to the environment.

Radiation in the Environment, Sources of. Table 26 US broken arrow accidents with environmental impact

No.	Date	Location	Weapons lost	Weapons destroyed
1	February 13, 1950	Coast off British Columbia, Canada		1
2	November 10, 1950	Rivière-du-Loup, Québec, Canada		1
3	March 10, 1956	Mediterranean Sea	2	
4	May 27, 1957	Kirtland Air Force Base, New Mexico, USA		1
5	July 28, 1957	Coast off New Jersey, USA	2	
6	January 31, 1958	Air Base, Sidi Slimane, French Morocco		1
7	February 5, 1958	Coast off Savannah, Georgia, USA	1	
8	July 6, 1959	Barksdale Air Force Base, Louisiana, USA		1
9	June 7, 1960	McGuire Air Force Base, New Jersey, USA		1
10	January 24, 1961	Goldsboro, North Carolina, USA		1
11	June 4, 1962	Johnson Atoll, Pacific Ocean		1
12	June 20, 1962	Johnson Atoll, Pacific Ocean	1	
13	December 8, 1964	Grissom Air Force Base, Indiana, USA		5
14	December 5, 1965	Ryukyu Islands, Japan	1	
15	January 17, 1966	Palomares, Spain		2
16	January 21, 1968	Thule Air Force Base, Greenland		4[a]
TOTAL			7	19

[a] Official US Air Force position is that fragments of all four nuclear weapons were found at Thule. There has been conjecture that one of the bombs melted through the ice and was lost on the sea bottom. An abundance of aircraft fragment was confirmed to be on the sea bottom. There is no readily available evidence to support the lost weapon hypothesis.

A contaminated area of 640 acres of hillside, village, and farmland was determined, of which 319 acres were cultivated farmland [84]. An aerial overview of a typical soil cleanup operation, depicting plow tractors, water dousing vehicles, and dump trucks removing soil, is provided in Fig. 23. Decontamination and cleanup criteria were based upon instrument readings which related to surface contamination of plutonium (Pu-239). Contaminated areas were assessed using portable alpha detectors (PAC-1S) as depicted in Fig. 24. An agreement was reached between governments of the USA and Spain that above a plutonium contamination level of 462 μg m^{-2}, soil was to be lifted, packaged, and transported from Spain to the USA. For lower levels of surface contamination, plowing and/or watering was deemed sufficient to mitigate the impact of the residual plutonium [85]. Contaminated soil, vegetation, and hot spots were packed into 4,810 barrels (Fig. 25) and shipped to Savannah River Nuclear Processing Facility for storage. Contaminated vegetation below an estimated 462 μg m^{-2} was removed to a dry riverbed and burned.

Environmental monitoring of the plutonium contamination in and around Palomares started immediately after the accident and continues to this day. Air monitoring (e.g., see Fig. 26) and soil and water sampling are performed on a continuing basis, as well as monitoring of local foodstuffs. Studies have shown persistent plutonium contamination in soils near the locations where weapons were destroyed [86]. Assessments of public doses from the event have been performed [87] and the estimates indicate that no member of the public would exceed the International Commission on Radiological Protection guidance level of 1 mSv a^{-1}.

Nuclear-Powered Vessel Accidents Nuclear powered vessels are floating, compact sources of large quantities

Radiation in the Environment, Sources of. Figure 23
Aerial view of soil cleanup site (Photo courtesy of Capt. Lewis B. Melson, US Navy (Ret))

Radiation in the Environment, Sources of. Figure 24
Airman conducting soil contamination survey (Photo courtesy of Capt. Lewis B. Melson, US Navy (Ret))

Radiation in the Environment, Sources of. Figure 25
Waste barrels with contaminated soil awaiting shipment to USA (Photo courtesy of Sandia National Laboratory)

Radiation in the Environment, Sources of. Figure 26
Chain link fence enclosing an air sampling station in Palomares (circa 2007) (Photo courtesy of Barbara Moran)

of fixed and loose radioactive material. In addition, unlike land-based commercial nuclear power plant counterparts, their operation is shrouded in secrecy. Many events, including the details of nuclear weapons as part of the vessels' armament inventory, are generally unavailable or speculative. As a result, the full magnitude of environmental releases of radionuclides from nuclear powered vessel accidents and incidents will never be known. It is known that a certain amount of radioactive material, from gases to coolant to resins, were (and possibly are) routinely released to marine environments during operation of these vessels. Also, nuclear-powered vessels often have tender ships that are utilized for, amongst other duties, reception of radioactive waste materials, which may lead to leakage to the environment under abnormal transfer. As a final environmental legacy of nuclear-powered vessel operations, it is well known that the former Soviet Union routinely disposed of decommissioned nuclear vessel components and reactors at sea, making the future environmental impact difficult to predict.

A list of known nuclear submarine accidents (all involving either the USA or USSR/Russian Federation) with known or potential environmental impact is provided in Table 27. It is quite possible that there are other accidents or incidents not included in this table that were never reported or remain classified to this day. A similar table for surface vessels is provided in Table 28. It is worthy to reiterate that under normal operations for these vessels, small routine releases occur.

The two principal threats to the environment from sunken nuclear vessels are (1) fission products in reactor fuel and (2) weapons material (primarily plutonium). It is worthy to note, however, that despite

Radiation in the Environment, Sources of. Table 27 Significant nuclear submarine with environmental impact

No.	Date and vessel	Location	Nuclear weapons lost	Reactor or fuel sunk	Radiation discharge
Submarines					
1	April 18, 1959, SSN-575 USS Sea Wolf	190 km east of Maryland, Atlantic Ocean		1 (scuttled reactor)	Possible
2	July 4, 1961, K-19 "*Hiroshima*"	South of Greenland, North Atlantic			Probable
3	April 10, 1963, SSN-593 USS Thresher	50 km east of Cape Cod, Massachusetts		1 (sunk)	Probable
4	May 22, 1968, SSN-589 USS Scorpion	740 km southwest of Azores, Portugal	2	1 (sunk)	Probable
5	March 8, 1968, K-129 Golf-II (NATO)	1,390 km northwest of Oahu, Hawaii	3 (sunk)		Possible
6a	May 24, 1968, K-27 "*Project 645*"	Reactor problem; location unknown			Probable
6b	September 6, 1982, K-27 "*Project 645*"	Scuttled in Kara Sea, north of Siberia		1 (scuttled vessel)	Possible
7	April 11, 1970, K-8 "*Project 627*"	490 km northwest of Spain, Bay of Biscay	4	2 (sunk)	Probable
8	October 6, 1986, K-219 Yankee I (NATO)	1,090 km northwest of Bermuda	16	2 (sunk)	Possible
9	April 7, 1989, K-278 Komsomolets	Barents Sea, Arctic Ocean	2	1 (sunk)	Yes
10	August 12, 2000, K-141 Kursk	Barents Sea, Arctic Ocean		2 (sunk)	Probable
11	August 30, 2003, K-159 Kit (NATO)	Barents Sea, Arctic Ocean		1 (scuttled fuel)	Possible
TOTAL			27	12	

Radiation in the Environment, Sources of. Table 28 Significant nuclear surface vessel with environmental impact

No.	Date and vessel	Location	Vessel category	Radiation discharge
Surface vessels				
1a	February 1965, NS Lenin	Unknown; Probably at shipyard	Nuclear-powered icebreaker (civilian)	Possible
1b	1967, NS Lenin	Unknown; Probably at shipyard	Nuclear-powered icebreaker (civilian)	Possible
2	December 12, 1971, AS-11 USS Fulton	Thames River, New Haven, Connecticut	Submarine tender	Yes
3	October 1975, AS-19 USS Proteus	Apra Harbor, Guam	Submarine tender	Yes

the moderate number of nuclear vessel accidents, no significant marine environmental contamination has been measured from these events. This is, in part, due to the depth at which these vessels sink or are scuttled, typically many thousands of meters, the large dispersion volume the oceans allow, and the complex pathway from release to receptor. It is likely that nuclear vessel reactors and weapons sunk or scuttled to great depths will remain in their sunken locations for as long as their local environment remains relatively static, although there are no universal models for evaluating either the short- or long-term impact of these events [88].

Depleted Uranium Munitions Depleted uranium (DU) is a mix of naturally occurring uranium isotopes in which the fissionable isotope U-235 is depleted compared to its abundance in natural uranium, as presented in Table 29.

DU has a high mass density (approximately 19 g cm^{-3}) and corresponding high electron density, which makes it ideal for radiation shielding, or any application that requires a large mass in a small volume. There are numerous uses for depleted uranium including radiation shielding, field trimmers for radiotherapy machines and aircraft control surface counterweights [89]. Therefore, on any aircraft crash where there is high velocity impact and/or fire, it is possible to get environmental contamination from DU. However, these are rare events and relatively small masses of DU, and are therefore of minimal environmental concern.

Radiation in the Environment, Sources of. Table 29 Percentage (by weight) of typical natural and depleted uranium

Isotope	Natural uranium (%)	Depleted uranium (%)
U-238	99.3	99.8
U-236		0.0003
U-235	0.72	0.2
U-234	0.006	0.0006

The utilization of depleted uranium in munitions and the subsequent dispersal into the environment is not an accident. However, the controversy surrounding the use of these munitions and their potential effect of both humans and the environment was not fully realized until the Persian Gulf War (“Gulf War I,” 1990–1991). Subsequently, DU ammunition was utilized extensively during the wars in Bosnia-Herzegovina (1992–1995), Kosovo (1999), Afghanistan (2002), and Iraq (“Gulf War II,” 2003).

Depleted uranium (DU) is a by-product of the enrichment of natural uranium process. It has the same chemical properties as natural uranium with approximately one third of the radioactivity. Its high density, pyrophoric properties, and low cost make it an ideal metal for military applications, wherein it is used for armor piercing ammunition. A typical DU penetrator (MK149-2) that is used, for example, in a Phalanx Close-in-Weapon System (CIWS), is shown in Fig. 27. This ammunition has a total length of approximately 5.4 cm (including a 1.5 cm plastic tip used to increase the stability of the ammunition in flight [90]), maximum diameter of approximately 1.2 cm with a mass of approximately 70 g. Other DU weapon systems, such as those utilizing anti-tank penetrators, are much larger (and have a much larger mass of DU) than the ammunition depicted in Fig. 27.

Other common DU ammunition types used by the US military are listed in Table 30. It is worthy to note that DU munitions are or have been used by a variety of countries other than the USA, and therefore the introduction of these sources into the environment either by military tactical use or in training is a worldwide issue.

When DU penetrators impact a hard target surface, the resultant high temperature causes small DU particulate to ignite both outside and inside the target. When the weapon penetrates the target, it generates shards and spall that can incapacitate the crew. As a result, a large amount of DU dust in the respirable range (<10 micron AMAD) is generated and dispersed. The DU fragments can be incorporated into humans through the shrapnel process, and/or DU dust may be inhaled by personnel in the vicinity of ammunition impact. Dust that is deposited in the environment can become mixed with surface layers of soil or sand, and can also become resuspended via wind or mechanical mixing. Once mixed with soils, DU will be sequestered in plants, and may enter the food chain this way (as natural uranium will) or may enter potable water supplies. When DU penetrators miss their targets, they embed into the ground and can contaminate the local surroundings [91].

Radiation in the Environment, Sources of. Figure 27 Depleted uranium armor piercing shell (Photo courtesy of E. Waller, UOIT)

Extensive environmental studies on air, soil, and water have been performed in military conflict regions in which DU has been used. Generally speaking, no detectable widespread contamination has been found in these regions. Contamination from DU tends to be limited to areas below penetrators that have missed targets and impacted soft soil or sand [92]. It has been suggested that the long-term effects of military use of DU on both local populations and the environment is not likely to be significant from either a chemical or radiological toxicity perspective [93].

Space Nuclear Power Source Accidents A historical and potential source of anthropogenic radioactive material being introduced into the environment is from the destruction on launch or reentry of a spacecraft utilizing a nuclear power source.

Radiation in the Environment, Sources of. Table 30 Depleted uranium ammunitions

Ammunition designation	Caliber (mm)	DU weight (g)	Used by
M829A1,A2	120	5,350	Tanks (M1 Abrams)
M900	105	4,246	Tanks (M1 Abrams)
PGU-14	30	298	Aircraft (A-10 Warthog)
M919	25	97	Armored personnel carriers (Bradley, LAV)
PGU-20	25	148	Heavy machine gun (MK-38)
			Aircraft (AV-8 Harrier)
MK149-2	20	70	Missile defense guns (Phalanx CIWS)

The two types of nuclear power sources that have been used on spacecraft are nuclear fission reactors and radioisotopic thermoelectric generators (RTGs). The mechanisms of operation vary greatly between the two. A space nuclear fission reactor works similarly to a commercial nuclear reactor, whereby the fission process generates heat, which may be used through a heat exchanger to run a turbine and generator (dynamic generation), or coupled to a transducer (similar to a thermocouple), which directly converts heat to electricity (static generation). An RTG, on the other hand, utilizes the properties of radioactive decay to generate heat which can be directly converted through a transducer to electricity (static generation). The advantage of both sources is that they can provide reliable and maintenance-free energy to power electronic systems in spacecraft, which is especially important in unmanned missions. Radiologically, in the case of a nuclear fission reactor (typically based on U-235), radioactive fission products are produced, as well as neutron activation of some surrounding materials. In the case of an RTG, the power source itself is highly radioactive. In both cases, destruction of the power source either at launch or through reentry can disperse radioisotopes over a large area, depending upon the position in the atmosphere where the breakup occurs.

To date, there have been eight verified accidents and two possible accidents involving space nuclear power source systems [94, 95]. All accidents have involved either USA or USSR (Russian after 1991) spacecraft, and are outlined in Table 31.

Of the few accidents involving space nuclear power systems, the two accidents known to have widely spread radionuclides into the environment are the Transit-5BN-3 and Cosmos-954 reentries, discussed below.

Transit-5BN-3 Reentry On April 21, 1964, the US Transit-5BN-3 navigation satellite was launched from Vandenburg Air Force base in California, but the rocket failed to boost the satellite into orbit. As a result, the payload reentered the southern atmosphere and the SNAP-9A RTG burned up [96]. The SNAP-9A RTG contained approximately 629 TBq of Pu-238, which is believed to have mainly deposited on the Earth's

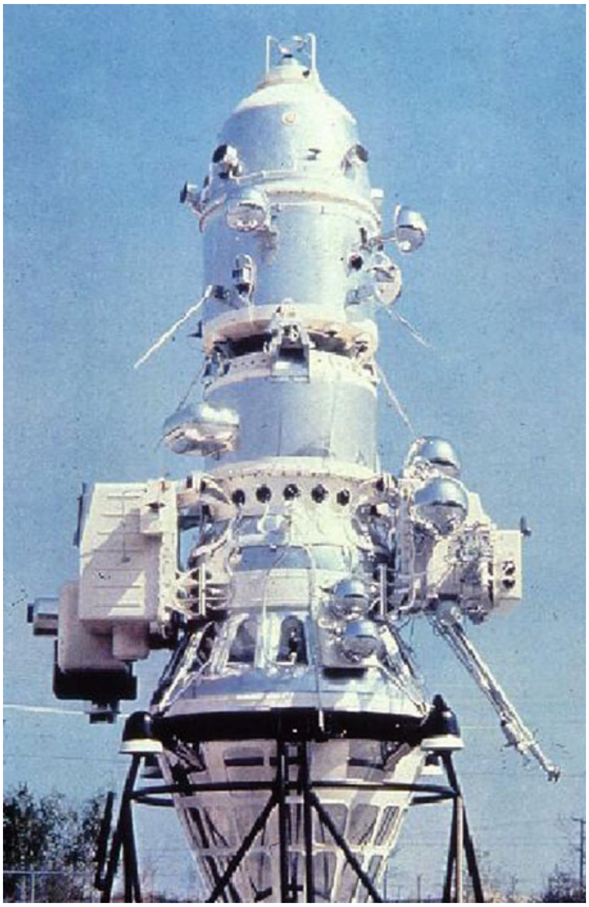
Radiation in the Environment, Sources of. Table 31 Space nuclear power system accidents

Accident date	Launch country	Spacecraft	Power source	Release to environment
Verified accidents				
April 21, 1964	USA	Transit-5BN-3	RTG (Pu-238)	Yes (see discussion below)
May 18, 1968	USA	Nimbus-1	RTG (Pu-238)	No (RTGs retrieved)
April 17, 1970	USA	Apollo 13	RTG (Pu-238)	No (RTGs never retrieved from ocean)
April 25, 1973	USSR	RORSAT	Unknown; probable nuclear reactor	Unknown (possible)
January 24, 1978	USSR	Cosmos 954	Nuclear Reactor (U-235)	Yes (see discussion below)
February 7, 1983	USSR	Cosmos 1402	Nuclear Reactor (U-235)	Unknown (possible)
September 30, 1988	USSR	Cosmos 1900	Nuclear Reactor (U-235)	No (reactor boosted to higher orbit)
November 16, 1996	Russia	Mars-96	RTG (Po-210)	Unknown (RTGs never recovered)
Possible accidents				
January 25, 1969	USSR	Unknown, possible RORSAT	Unknown; probable nuclear reactor	Unknown
September 23 and October 22, 1969	USSR	Unknown; unmanned moon probes	Unknown; possible Po-210 RTG source	Probable

surface [94]. This deposition was estimated to have almost tripled the global deposition of Pu-238 by 1970 (the prior contribution being from global nuclear weapons testing).

Cosmos-954 The USSR Cosmos-954 (also known as Kosmos-954) was launched on September 18, 1977 (Fig. 28). The satellite was postulated to be powered by an approximately 20 kg U-235 nuclear fission reactor.

On January 24, 1978, the satellite reentered the atmosphere with the nuclear reactor core on-board, partially burning up over a 1,000 km long path extending primarily over the Canadian Northwest Territories, and partially into the provinces of Alberta and Saskatchewan. A large-scale search was conducted for potential contamination over approximately 124,000 km^2 extending from January to October 1978. In Fig. 29, a radiation surveyor is seen performing duties in the harsh Arctic environment, which slowed recovery efforts. Figure 30 shows two persons packaging radioactive material for removal from recovery site.



Radiation in the Environment, Sources of. Figure 28 Cosmos satellite (Adapted from http://gsc.nrcan.gc.ca. Reproduced with the permission of the Minister of Public Works and Government Services, 2010)

At the time of reentry, the core was estimated to contain approximately 3.1 TBq Sr-90, 180 TBq I-131, and 3.2 TBq Cs-137 [25]. The reentry generated a dispersion of radioactive fragments and particles from large chunks (Fig. 31) to particles less than 100 μm in diameter (Fig. 32). The fragments depicted in Fig. 31 were termed the “antlers,” as from a distance they were originally thought to be Caribou antlers. The “antlers” were believed to be part of the reactor control mechanism [97]. An additional large piece of debris, termed the “stovepipe” due to its shape (Fig. 33), was found to be nonradioactive and likely the only material found that was not part of the reactor structure. Numerous beryllium rods were discovered (Fig. 34), which were hypothesized to be part of the neutron reflector system [97].

The Cosmos 954 satellite is believed to have weighed 4–5 t; however, only 65 kg total mass of material was recovered [97]. It is estimated that approximately 75% of the original material remained in the upper atmosphere. Approximately 0.1% of the particulate dispersed is estimated to have been recovered by the search efforts. Most of the recovered material was radioactive, suggesting that it was part of the reactor structure. A summary of the recovered debris is provided in Table 32 (adapted from [97]).

The most radioactive fragment (Fig. 35), which was most certainly a piece of the reactor core, was found to have a dose rate of approximately 5 Gy h^{-1} near contact.

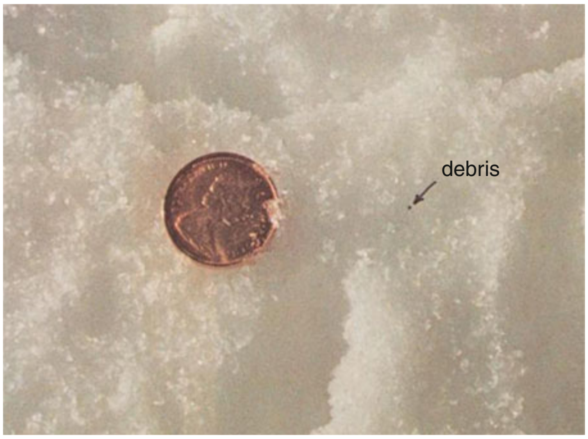
Analysis of the fragments and particles identified a number of fission and activation products characteristic of an enriched U-235 reactor (Table 33).

Based upon the difficulty of finding small radioactive particles dispersed over a large area, it is highly probable that radioactive particles, and perhaps fragments, from the reentry of Cosmos-954 remain in Northern Canada to this day as environmental contamination. From the list outlined in Table 33, it is likely that the principal isotopes detectable over

Radiation in the Environment, Sources of. Figure 29
Radiation surveyor approaching piece of Cosmos 954 debris (Photo courtesy of Canadian Nuclear Safety Commission [97])

Radiation in the Environment, Sources of. Figure 30
Radiation surveyors packaging debris from Cosmos 954 (From http://gsc.nrcan.gc.ca. Reproduced with the permission of the Minister of Public Works and Government Services, 2010)

Radiation in the Environment, Sources of. Figure 31
Large piece of satellite debris nicknamed the "antlers" (Reproduced from www.nsarchive.org with the permission of the National Security Archive)



Radiation in the Environment, Sources of. Figure 32
Small radioactive particle of fuel (Photo courtesy of Canadian Nuclear Safety Commission [97])

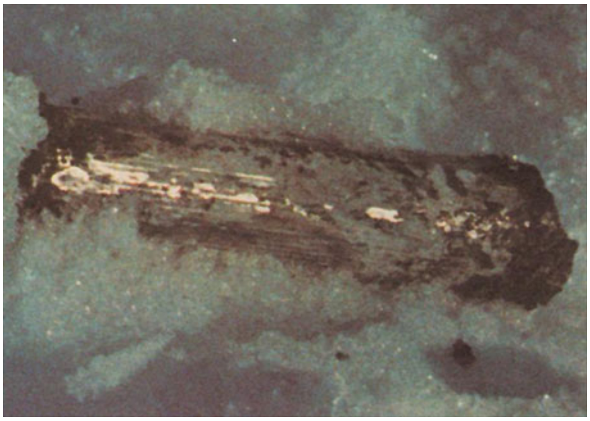
Radiation in the Environment, Sources of. Figure 33
The "stovepipe" debris (Photo courtesy of Canadian Nuclear Safety Commission [97])

30 years post crash would be H-3, Cs-137, Sr-90, and Pu-239. Of the four principal radionuclides identified, the only one that would be readily detectable in a field survey due to its gamma emission would be Cs-137.

Medical, Industrial, Commercial, and Research Sources

There are numerous types and categories of radioactive materials used for medical, industrial, commercial, or research uses. The sources can be gaseous, liquid, or solid in form. Sources that are in gas or liquid form can directly enter the environment as they are inherently mobile. Sealed sources (generally in solid form) can only enter the environment if there is a disruption to their source capsule. On a basis of specific activity, sealed sources generally have the highest activity.

Figure 36 graphically depicts the range of source activity for common sealed radiation sources [98]. The activity of these sources varies from hundreds of Bq to the PBq range.



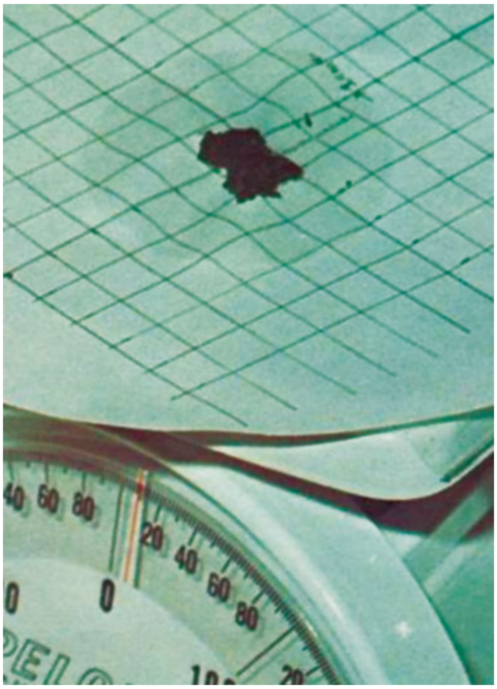
Radiation in the Environment, Sources of. Figure 34 Beryllium rod found in snow (Photo courtesy of Canadian Nuclear Safety Commission [97])

Radiation in the Environment, Sources of. Table 32 Summary of recovered debris from Cosmos 954

Item(s)	Mass (g)	Max. Measured dose rate (Gy h^{-1}) on contact
Large Debris		
Steel plate fragments (x4)	272 (max)	2
Beryllium rods (x41)	51 each	1
Beryllium cylinders (x6)	3,600 each	0.15
Tubes/rods/plate ("antlers")	20,000 total	0.15
Steel tube ("stovepipe")	18,200 total	Not detectable
Chunks/flakes/slivers	Variable	5
Small Debris		
~4,000 particles	1E-4 to 5E-3	1E-3

Details of these sources are briefly discussed below.

Teletherapy units are commonly found in medical institutions, such as hospitals or clinics. The physical dimensions of the source are relatively small, with generally a cylindrical (few centimeters in diameter by several centimeters long) shape. The source is contained inside a large shielding device.



Radiation in the Environment, Sources of. Figure 35 Most radioactive fragment found from Cosmos 954 (Photo courtesy of Canadian Nuclear Safety Commission [97])

Irradiator facilities are relatively few in number, and contain very high activity sources to sterilize foodstuffs, medical products and supplies, and for other specialized applications. The sources used in performing the irradiation of the material vary in physical size, some being large or others being pencil sized, and each facility will contain many such sources.

Portable industrial radiography sources and their devices are generally small in terms of physical size, although the devices are usually heavy due to the shielding contained in them. The sources themselves are very small, less than 1 cm in diameter, and only a few centimeters long, and are attached to specially designed cables for their proper operation. The use of radiography sources and devices is very common, and their portability may make them susceptible to theft or loss. The small size of the source allows for unauthorized removal by an individual, and such a source may be placed into a pocket of a garment. Industrial radiography may also be performed in fixed installations, either using the same small portable devices, or using larger machines which may appear to be similar to teletherapy units.

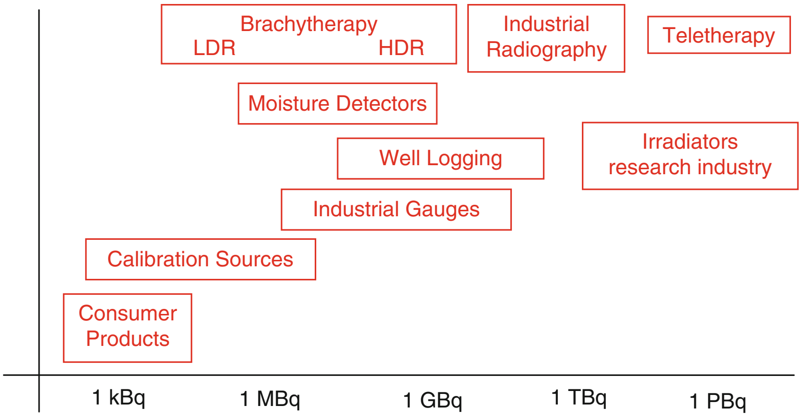
Brachytherapy sources are of two slightly different varieties. These are generally referred to as Low Dose Rate (LDR) brachytherapy and High Dose Rate (HDR) brachytherapy. Both applications use sources that may be small physically (less than 1 cm in diameter, only a few cm long) and, thus, are susceptible to being lost or misplaced. HDR sources, and some LDR sources, may be in the form of a long wire attached to a device (a remote afterloading device). The afterloading device may be heavy, due to the shielding for the sources when not in use, and the device may be on wheels for transport within a facility. The remote afterloading device may also contain electrical and electronic components for its operation, and pneumatic systems for source transfer. Brachytherapy sources are located in hospitals, clinics, and similar medical institutions, and such facilities may have a large number of sources.

Well logging sources and devices are generally found in areas where exploration for minerals is occurring, such as searching for coal, oil, natural gas, or similar uses. The sources are usually contained in long (1–2 m, typically) but thin (<10 cm in diameter) devices which also contain detectors and various electronic components. The actual size of the sources inside the devices is generally small. The devices are heavy, due to the ruggedness needed for the environments in which they are used.

Industrial sources (gauges) are of various shapes and sizes, and are either fixed or portable. These devices are generally designed for many years of operation

Radiation in the Environment, Sources of. Table 33
Fission and activation products detected in Cosmos 954 debris

Fission product	Activation product
^{89}Sr	^{51}Cr
^{90}Sr	^{54}Mn
^{95}Zr	^{58}Co
^{95}Nb	^{60}Co
^{99}Mo	^{59}Fe
^{137}Cs	^{182}Ta
^{140}Ba	^{46}Sc
^{140}La	^{124}Sb
^{141}Ce	^{239}Pu
^{144}Ce	^{3}H
^{103}Ru	
^{106}Ru	
^{132}Te	
^{131}I	
^{147}Nd	



Radiation in the Environment, Sources of. Figure 36
Activity range of common sealed sources used in medicine and industry

with little or no special tending. They may be used for control over a process, for measurement of flow, volume, density, material presence, and may be placed in locations unsuitable for continuous human presence. Consequently, they may accumulate layers of dirt, grime, grease, oil, material, etc., covering any warning labels that may have been present. Depending upon the specific application, industrial gauges may contain relatively small quantities of radioactive material, or may contain sources with activities approaching 1 TBq.

Moisture/density devices are types of industrial gauges which are small and portable. These devices contain the sources, detectors, and electronic gear necessary for the measurement undertaken. The source is physically small in size, typically a few cm long by a few cm in diameter, and may be located either completely within the device or at the end of a rod/handle assembly.

Calibration sources are typically small sealed sources that are used to calibration and verification of the operation of radiation detection equipment.

Consumer products include any commercially available item that utilizes a radionuclide source to operate. A common example is a smoke detector.

Table 34 lists a number of sealed source categories by application, radioisotope, and typical radioactivity level [99]. The source categories (Category 1, 2, and 3) are for those requiring the most stringent control (Category 1) to the least (Category 3) as determined by the IAEA [100]. There is generally no pathway to the

Radiation in the Environment, Sources of. Table 34 Common sources, categories, and activities

Application	Radioisotope	Typical radioactivity (GBq)
Category 1		
Radioisotope thermoelectric generators	Sr-90	1E+06 – 1E+07
Teletherapy	Co-60	5E+04 – 1E+06
	Cs-137	5E+05
Blood irradiation	Cs-137	2E+03 – 1E+05
Industrial radiography	Ir-192	1E+02 – 9E+03
	Co-60	1E+02 – 9E+03
Sterilization and food preservation (irradiators)	Co-60	1E+05 – 4E+08
	Cs-137	1E+05 – 4E+08
Other irradiators	Co-60	1E+03 – 1E+06
	(Cs-137 rare)	1E+03 – 1E+06
Category 2		
High dose rate remote afterloading brachytherapy	Co-60	1E+01
	Cs-137	3E-05 – 1E-02
	Ir-192	4E+02
Low dose rate brachytherapy (manual or remote)	Cs-137	5E-02 – 5E-01
	Ra-226	3E-02 – 3E-01
	Co-60	5E-02 – 5E-01
	Sr-90	5E-02 – 1E+00
	Pd-103	5E-02 – 1E+00
Well logging	Cs-137	1E+00 – 1E+02
	Am-241/Be	1E+00 – 8E+02
	Cf-252 (rare)	5E+01
Level gauge	Cs-137	1E+01 – 1E+03
Thickness gauge	Co-60	1E+00 – 1E+01
Conveyor gauge	Am-241	1E+01 – 4E+01
Moisture/density detector (portable, mobile units)	Am-241/Be	1E-01 – 2E+00
	Cs-137	≤ 4E-02
	Ra-226/Be	1E+00
	Cf-252 (rare)	3E+00
Category 3		
Level gauge	Cs-137	1E-01 – 2E+01
Density gauge	Co-60	1E-01 – 1E+00
Thickness gauge	Kr-85	1E-01 – 3E+00
	Am-241	1E+00 – 1E+01
	Sr-90	1E-01 – 4E+01
	Tl-204	4E+0

environment for these sealed source categories unless they are (a) destroyed, (b) tampered with, or (c) intentionally dispersed maliciously.

Medical Sources Introduction of large sealed medical sources into the environment from, for example, teletherapy units would generally only occur through complete destruction of the source via fire or explosion, or theft of the source and inadvertent or intentional dispersion. Such a scenario is considered in section on "Lost and Orphan Sources" and is not further considered here.

Medical sources of radiation are routinely introduced into the environment from nuclear medicine procedures. Isotopes utilized in nuclear medicine diagnostic imaging are characterized by their short radiological half-lives and relatively low energies of gamma emission. Common sources used in nuclear medicine are presented in Table 35. A more detailed discussion of nuclear medicine pharmacy can be found in [101].

Radionuclides can be used in pure form, or attached with some other chemical molecule to perform a specific body function (termed "labeling" or "radiolabeling"). Radionuclides used in medicine are called radiopharmaceuticals, and consists of two components: the radionuclide and the pharmaceutical. The design of a radiopharmaceutical is based upon two characteristics: (1) its ability to preferentially locate in a desired organ or for its participation in a physiological function, and (2) the emission of radiation that will either be detected (for imaging) or deposit energy (for treatment). Approximately 95% of radiopharmaceuticals are used for diagnostic imaging, while the remaining 5% are used for therapy. Of the diagnostic imaging radiopharmaceuticals, approximately 80% are Tc-99m-labeled compounds [101].

Technetium has become the most widely used radionuclide for diagnostic nuclear medicine. It has advantageous physical characteristics of short half-life, low energy of its mono-energetic gamma ray and ease of chelation. It is formed from the decay of a parent radionuclide, molybdenum-99, which through this parent–daughter process, can be provided in a convenient, readily available and mobile form, the Technetium Generator. Tc-99m is formed when molybdenum-99 (^{99}Mo, popularly called "Molly") emits radioactivity. Molly has a half-life of 66 h and is easily bound to the cartridge. As Molly gradually disintegrates, ^{99m}Tc is formed which does not bind as well to the cartridge and can be washed out as needed with sterile salt water. The cartridge is delivered to the customers in a lead container which protects against radiation exposure. These are called technetium generators, or "cows," as they can be "milked" repeatedly as needed. The technetium is combined with a "label" and injected into the patient. After a waiting period to allow the Tc-99m-labeled compound to be distributed to target organs, the patient is scanned with an imaging device. Almost all of the technetium leaves the body via urine or radiologically decays within the body.

Radiation in the Environment, Sources of. Table 35 Some common radionuclides used in nuclear medicine studies

Element	Radionuclide	Half-life	Energy gamma (keV)	Energy beta (keV)
Phosphorus	^{32}P	14.3 days		1,700
Chromium	^{51}Cr	27.7 days	320	
Gallium	^{67}Ga	78.3 h	93, 184, 300	
Strontium	^{89}Sr	50.5 days		1,480
Yttrium	^{90}Y	64.1 h		2,280
Technetium	^{99m}Tc	6.02 h	140	
Indium	^{111}In	2.8 days	171, 245	
Iodine	^{131}I	8.0 days	364	606
	^{123}I	13.2 h	159, 35	
Thallium	^{201}Tl	73.1 h	68–82 (x-rays) 167	

The Tc-99m that is excreted via urine enters the municipal sewer system and ultimately enters the aquatic environment. Due to the short radiological half-life, low energy of the decay emissions, and the fact that hospital delivery is regulated to acceptable release levels, environmental impact of Tc-99m on the environment is minimal.

Another nuclear medicine radionuclide that is sometimes observed in the environment is I-131. Radioactive iodine can be observed in aquatic plants near harbors where hospitals perform thyroid imaging or treatment using radioactive iodine [76]. Although the radiological half-life of I-131 is only 8 days, a constant introduction into the environment can yield a steady state burden of I-131 in aquatic plants. In addition, I-131 as well as other nuclear diagnostic radionuclides have been observed in river sediment downstream from hospitals performing nuclear medicine procedures [102]. It is noted that radionuclides in the environment from medical procedures far exceed equivalent releases from nuclear power plants [102]. No known deleterious effects of these radionuclides have been observed in either human or nonhuman biota, primarily due to the fact that their release is regulated to levels below regulatory concern. These radionuclides do, however, make up part of the food chain for both human and nonhuman biota and must be taken into account in total dose estimates.

Industrial Sources Although there are various industrial applications of radioactive sources, the most common classifications are:

1. Radiography sources
2. Nuclear gauges
3. Irradiator sources
4. Radioactive tracers

Except for tracers, all of the above categories of sources are sealed activity, and therefore their introduction into the environment requires damage or intentional removal of the source material from its protecting capsule. The sealed sources are briefly discussed below.

Industrial radiography sources come in a wide variety of packaging. Some are in the form of cameras (which work on a shutter mechanism) and some are in the form of a projector Teleflex cable. Industrial radiography requires relatively high activity sources that emit energetic gamma rays. Examples of radioisotopes commonly used in radiography include ^{60}Co, ^{137}Cs, ^{192}Ir, and ^{169}Yb. The source for a projector-type industrial radiography device is typically located at the end of a Teleflex pigtail. The pigtails have quick-release connectors for ease of source replacement. Typically, the source is encapsulated within two welded stainless steel capsules. The source material is often radioactive metal (cobalt or iridium), or embedded on nonsoluble ceramics (called microspheres) that minimize the potential of inhalation if the source encapsulation is breached. For transportation from source manufacturer to end user, the capsule is typically positioned at the center of a drum, surrounded by lead and shock-absorbent packing.

The basic principle of a nuclear gauge is that a radiation source is placed on one side of a material or container that requires a measurement (such as mass density, level of material, etc.), and a detector is placed on the other side. For example, a source is placed on one side of a pipe, and the detector on the other side. The amount of attenuation of the primary beam is proportional to the material between the source and detector. Therefore, this type of arrangement is useful for mass flow rate or mass density measurements. The source for this type of application is generally a photon source on the order of tens of gigabecquerel. Another popular type of gauge is the Troxler^{TM} density and moisture gauge (Fig. 37). The Troxler device uses two types of radiation sources: gamma (~296 MBq ^{137}Cs) for measuring density and neutron (~1.5 GBq ^{241}Am-Be or ~2.2 MBq ^{252}Cf) for measuring moisture content.

An irradiator is a radiation device containing radioactive material, usually ^{137}Cs or ^{60}Co, in the form of a sealed source. The activity of the source is relatively high, often ten to hundreds of petabecquerel. These sealed sources are used to deliver very high radiation doses to objects and biological materials for various purposes. For example, irradiators are used to (a) sterilize blood, soil, medical instruments, and food; (b) to calibrate radiological instruments; (c) to extend the potency of cells; (d) in the treatment/eradication of cancer; or (e) to measure the density and detect deficiencies of various materials. There are two basic categories of irradiators: self-contained (enclosed beam) and non-self-contained (open beam or panoramic)

Radiation in the Environment, Sources of. Figure 37 Troxler^{TM} density and moisture gauge (Photo courtesy of E. Waller, UOIT)

irradiators. Examples of open-beam irradiators are teletherapy, large wet-source-storage (pool-type), and panoramic dry-source-storage. Self-contained irradiators are devices in which the primary radiation beam is completely shielded during use and storage conditions. Consequently, under normal conditions of use, there are no high levels of radiation within the irradiator room. An example of a self-contained irradiator is the MDS Nordion Gammacell 220. The unit is 193 cm high, 107 cm diameter, and mass of approximately 5,680 kg. The sample is placed in a chamber 20 cm high by 15 cm diameter and lowered into the cell. This type of irradiator uses ^{60}Co, and the activity (new) can range from 89 to 890 TBq. The central dose rate delivered to a sample ranges (based on activity) from approximately 2 to 20 kGy h^{-1}. The dose rate at 5 cm from the outer surface of this irradiator is on the order of hundreds of μSv h^{-1}. Non-self-contained or panoramic irradiators are devices in which the primary radiation beam is not shielded during irradiator use. It is shielded only when the source is in storage position. The type of shielding used depends on the type of irradiator. For example, water in the form of a storage pool is the shielding material for pool-type or wet storage irradiators whereas lead and concrete are often used for dry-source-storage and teletherapy-type irradiators. During use, the source is not shielded and there are very high radiation levels in the irradiator room. In a typical irradiation facility used for sterilization of medical supplies, cartons are moved to a central location, and high activity gamma sources are raised into an irradiate position. An example is the MDS Nordion JS 10000 hanging-tote irradiator. This irradiator uses up to 1.85 PBq ^{60}Co sources in the form of racked pencils, and is capable of delivering tens to hundreds of thousands of gray per hour, depending upon the configuration.

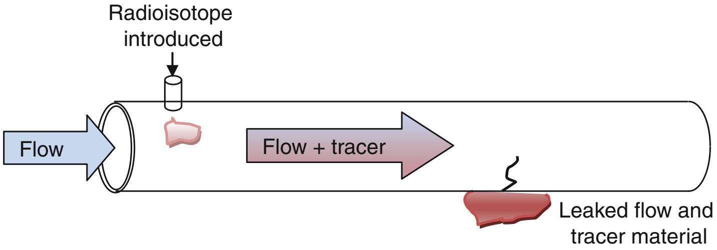
Radionuclides are used as tracers since they emit a characteristic radiation signature at very low material mass. Tracers are useful when transport of material is being characterized in terms of flow velocity, direction, and interaction with the environment. Another extremely useful application of tracers is determination of a leak in a closed system. For example, consider a pipe carrying a liquid where it is suspected that there is a leak in the pipe. A radionuclide is introduced at an upstream position in the pipe and a radiation detection instrument is used to find an area of large signal, indicating an accumulation of flow material, which includes the radionuclide (Fig. 38).

Specific uses of radionuclide tracers include [7]:

- Chemical process investigations, such as flow rate and residence time determination
- Mixing time and optimization
- Maintenance investigations, such as leak detection
- Wear and corrosion investigations, including lubrication studies

Generally speaking, industrial sources are only an environmental issue when they are disrupted, as non-sealed source introduction into the environment is done with very low level of activity using short-lived isotopes.

Commercial Sources There are various applications of radionuclides for commercial purposes. For example, historically, radium (Ra-226) was used in self-luminous light sources. Today, tritium (H-3) is found for the same application in watches, emergency lighting, and any location where a light indicator is required without a power source. Some consumer products containing radionuclides that may find their way into the environment due to the principal fact that they are uncontrolled substances once released to the consumer, are presented in Table 36 [103]. Modern consumer



Radiation in the Environment, Sources of. Figure 38
Application of tracer for leak detection

Radiation in the Environment, Sources of. Table 36 Consumer items containing radionuclides

Consumer product	Radionuclide(s)	Comment
Smoke detectors	^{241}Am	In older smoke detectors, ^{90}Sr was also used.
Watches/clocks	^{3}H	Historically (pre-1970s), ^{226}Ra was the predominant radionuclide used.
Ceramics	Uranium chain	Older (pre-1960s) ceramics can contain significant quantities of uranium and progeny.
Glassware	Uranium chain	Historical glass ("Vaseline glass") often contained uranium for coloring.
Lenses	Thorium chain	Some lenses produced from 1950s through 1970s used ^{232}Th as a coating.
Lantern mantles	Thorium chain	Historically, many gas lantern mantles contained significant quantities of ^{232}Th.
Food	Natural radionuclides	Food is a commodity that inherently contains radionuclides, sequestered from the environment.
Fertilizer	^{40}K, uranium chain	Fertilizer is high in potassium naturally and can contain phosphates with elevated uranium concentrations.
Historical items	Primarily ^{226}Ra	Many historical items, "quack cures" and the like around the beginning of the twentieth century employed radioactivity for various purposes.

products utilizing radioactive material are highly regulated to ensure no appreciable dose to the public from either their use or disposal. Historical items and artifacts that may still be found in homes, antique shops, museums, or on Internet auction sites may still contain significant quantities of radioactive material (often as uranium/radium decay series). The historical items can become sources of radioactivity if disposed improperly.

The two most common consumer products containing anthropogenic radioactivity are described below.

Gaseous Tritium Light Sources Tritium may be obtained in relatively large quantities as gaseous tritium light sources (GTLS). These sources are commonly used in emergency lighting systems and watches (Fig. 39).

Tritium gas is readily oxidized or exchanges with other hydrogen isotopes forming HTO or tritium oxide which can remain in the atmosphere for a time or precipitate to the Earth [2]. However, given the relatively small risk associated with the ingestion of tritium (short biological half-life and weak beta emitter), it is not a significant environmental source of radioactivity in this form.

Smoke Detector Modern ionization-based smoke detectors use ^{241}Am (half-life ~432 years,) which decays by alpha emission. A typical detector uses an activity of approximately 37 kBq. The ^{241}Am source is located in an ionization chamber consisting of two plates with a voltage across them, and the source near one plate. The alpha particles ionize the oxygen and

Radiation in the Environment, Sources of. Figure 39
GTLS in wristwatch (Photo courtesy of E. Waller, UOIT)

nitrogen atoms of the air in the chamber. The electrons from the ionizations are attracted to the plate with a positive voltage, and the positively charged atom is attracted to the plate with a negative voltage plate. The electronics in the smoke detector sense the small amount of electrical current that these electrons and ions moving toward the plates represent. When smoke enters the ionization chamber, it disrupts this current – the smoke particles attach to the ions and neutralize them. The smoke detector senses the drop in current between the plates, and alarms accordingly.

Research Sources In university and research settings, virtually any category of radioactive source may be utilized, dependent upon the nature of the research. Radionuclides are used in medicine, biology, physics, chemistry, geology, and engineering studies. Very common research sources used in research settings are presented in Table 37 [104].

Laboratories and research activities are generally well regulated such that environmental releases and radioactive waste disposal are governed by the license conditions to hold the isotopes.

Lost and Orphan Sources

The IAEA (2002) estimates that millions of radioactive sources have been distributed globally over the past 50 years. In addition, the IAEA estimates over 20,000 operators of significant radiation sources, including radiotherapy units (more than 10,000), industrial radiography sources (more than 12,000) and approximately 300 irradiator facilities [105]. Given the vast quantity of sources distributed globally, the variable regulations and regulatory guidance over the past 50 years and regulatory differences from country to country, it is not implausible that some of these sources have become lost, stolen, or orphan.

Radiation in the Environment, Sources of. Table 37
Common research radionuclides

Radionuclide	Typical use
^{3}H	Biological tracer studies
^{14}C	Biological tracer studies
$^{32,33}P$	Biochemical labeling studies
$^{125,131}I$	Medical research
^{60}Co	Irradiation and physics studies
^{137}Cs	Physics studies
^{252}Cf	Neutron studies

A “lost source” is a source that has been misplaced, has become detached from a shielded unit, has been stolen, or otherwise has departed from positive control of a responsible party. The term “orphan source” refers to a sealed source of radioactive material contained in a small volume, but not radioactively contaminated soils and bulk metals, in any one or more of the following conditions:

- In an uncontrolled condition that requires removal to protect public health and safety from a radiological threat.
- Controlled or uncontrolled, but for which a responsible party cannot be readily identified.
- Controlled, but the material’s continued security cannot be assured. If held by a licensee, the licensee has few or no options for, or is incapable of providing for, the safe disposition of the material.
- In the possession of a person, not licensed to possess the material, who did not seek to possess the material.
- In the possession of a radiological protection program for the sole purpose of mitigating a radiological threat because the orphan source is in one of the conditions described in one of the first four

bullets, and for which the program does not have a means to provide for the material's appropriate disposition.

There have been a significant number of lost and orphan sources estimated worldwide. The USA estimates that 1,500 sources have been lost since 1996 with more than half never being recovered, and that currently approximately 375 sources per year are reported lost or stolen [106]. The European Union estimates up to 70 sources per year are lost from regulatory control [105]. Any one of these source, if tampered with, can introduce anthropogenic radionuclides into the environment which can result in harm to human and nonhuman biota, as well as cause significant psychological, social, and economic burden. A number of source incidents impacting the environment are discussed in the next section.

Specific Source Incidents The incidents discussed below do not represent all lost or orphan source accidents, although they do represent some major accidents and the potential consequence. Generally speaking, when a lost source is found, it is relatively straightforward to mitigate the effects. However, it is probable that there are many lost or orphan sources that may have breached source capsules and are leaking radioactivity into the environment.

- January 2002 (Tbilisi, Georgia) Two orphan ^{90}Sr sources (~1.11 PBq each) used as portable electrical generators known as "radioisotopic thermoelectric generators" were recovered from the forest near Tbilisi (Lja). The sources had been found by woodsmen, who used them to keep warm (all three suffered varying degrees of radiation injury). At least two similar Sr-90 units are believed to be missing in Georgia.
- January–February 2000 (Samut Prakan, Thailand) A number of teletherapy heads were in unauthorized storage in a warehouse. Thieves accessed the site and partially disassembled one unit containing a ^{60}Co source (15.7 TBq). The thieves removed the head to their residence and continued to disassemble, which resulted in the source capsule falling out of its housing unnoticed. A number of people received high doses, and three people died of radiation exposure.
- May 2000 (Cairo, Egypt) A ^{192}Ir (3 TBq) radiography source fell from the back of a truck transporting it, and a farmer and his son found it in a ditch. They took the item home and a number of family members were exposed to high doses of radiation. The farmer and his son died of radiation exposure.
- February 1999 (Yanango, Peru) An industrial radiographer failed to notice that his ^{192}Ir (1.17 TBq) had disconnected from its Teleflex cable and was lost in the area in which he had been working. Another worker picked up the object and removed it to his residence. The worker, his wife, and baby were exposed to the source. The worker received extremely high doses and subsequently lost his leg due to the exposure.
- December 1998–January 1999 (Istanbul, Turkey) Three used ^{60}Co teletherapy sources that were meant to be shipped back to point of origin in the USA, were left in storage for 5 years. Tracking on the packages was lost and two of the packages were subsequently sold for scrap metal. The scrap merchants opened the source container (~3.3 TBq activity) and were exposed. A number of people received doses from 0.1 and 3.0 Gy. The third source package of 23.5 TBq was never found.
- October 1997 (Lilo, Georgia) A large number of ^{137}Cs and ^{226}Ra sources, abandoned at a former Russian training center, were found in the clothing of Georgian military personnel at the Lilo Training Facility, and around the grounds and local environment. The sources, many of which were in close contact with the soldiers for extended periods of time, caused varying degrees of radiation injury to the personnel involved.
- 1983 (Juarez, Mexico) An orphaned ^{60}Co teletherapy source (37 GBq) from a hospital was sold as scrap metal. During transport some small source pellets scattered along the road, and the unit was subsequently smelted and used in building materials (rebar). A number of people received doses ranging from 0.25 to 7 Gy. A total of 814 homes were demolished, generating 16,000 m^3 of soil waste and 4,500 t of metal waste.

A highly publicized and dramatic example of the potential for environmental and human catastrophe from lost or orphan sources is the Goiânia accident, discussed below.

Goiânia Accident (1987) Possibly the most well-known orphan source case leading to injury, death, and environmental contamination is the Goiânia accident. In 1985, the Instituto Goiano de Radioterapia in Goiânia, Brazil, moved to new premises, taking with it a ^{60}Co teletherapy unit and leaving in place a ^{137}Cs teletherapy unit without notifying the licensing authority as required under the terms of the institute's license. The former premises were subsequently partly demolished and as a result, the ^{137}Cs teletherapy unit became unsecured. On September 13, 1987, two people entered the premises and, not knowing what the unit was but thinking it might have some scrap value, removed the source assembly from the radiation head of the machine and tried to dismantle it at home. In the attempt the source capsule was ruptured. The radioactive source was in the form of ^{137}CsCl (cesium chloride salt) which is highly soluble and readily dispersible. The source activity was 50.9 TBq, with a total source plus inert matrix mass of 93 g [107] and estimated ^{137}CsCl mass of 22 g. For scale, approximately 22 g of (nonradioactive) CsCl is depicted in the hand of the author in Fig. 40.

Contamination of the environment ensued with the external irradiation and internal contamination of several persons. After the source capsule was ruptured, the remnants of the source assembly were sold for scrap to a junkyard owner. Fragments of the source, the size of rice grains, were distributed to several families who were fascinated that the material glowed blue in the dark. This contact with the source continued for 5 days, by which time a number of people were showing gastrointestinal symptoms arising from their exposure to radiation from the source. The symptoms were not initially recognized as being due to irradiation. However, one of the persons irradiated connected the illnesses with the source capsule and took the remnants to the public health department in the city. This action began a chain of events which led to the discovery of the accident [107]. Many individuals incurred external and internal exposure. In total, approximately 112,000 persons were monitored, of whom 249 were contaminated either internally or externally, and 4 died from their exposures. Some suffered very high internal and external contamination owing to the way they had handled the ^{137}CsCl powder, such as daubing their skin and eating with contaminated hands, and via contamination of buildings, furnishings, fittings, and utensils. Massive amounts of resources were utilized both for the initial response, follow-up, and environmental remediation. Hundreds of personnel were required to perform initial radiological assessments on thousands of people and to remediate the city. The response also generated large quantities of radioactive waste. A temporary waste storage site was chosen 20 km from Goiânia. Wastes were classified into nonradioactive (below 74 kBq kg^{-1}), low level (below 2 mSv h^{-1}) and medium level (between 2 and 20 mSv h^{-1}). Various types of packaging were used, according to the levels of contamination. The waste packaging required 3,800 metal drums (200 L), 1,400 metal boxes (5 t), 10 shipping containers (32 m^{3}), and 6 sets of concrete packaging. The temporary storage site was designed for a volume of waste of 4,000–5,000 m^{3} encapsulated in about 12,500 drums and 1,470 boxes. The final total volume of waste stored was 3,500 m^{3}. The economic burden of such levels of waste is enormous [107], and long-term socioeconomic effects were devastating. Goiânia suffered a 20% drop in gross domestic product, which took 5 years to return to normal levels. Tourism in the tropical town dropped to zero and Goiânia found itself the victim of economic

Radiation in the Environment, Sources of. Figure 40
Simulated ^{137}CsCl powder source of mass 22 g in hand (Photo courtesy of E. Waller, UOIT)

discrimination, as demand for food and other products from the area plummeted. Even with careful monitoring and accounting, the estimate of activity in the recovered contaminated materials is approximately 44 TBq, whereas the activity of the ^{137}Cs as stored in the teletherapy head is known to be 50.9 TBq. This suggests that even with careful analysis, survey and recovery, it is very difficult to recover all of a widely distributed source. It is also worthy to note that the people in and around Goiânia were considered pariahs around Brazil. People were afraid both of the people from the area and the area itself. The local fear and anxiety generated by this event was widespread and lasting.

Radiological Dispersal Devices

A radiological dispersal device (RDD), sometimes referred to as a "dirty bomb," is a terrorist weapon. Terrorism may be defined as the systematic use of terror or unpredictable violence against governments, public, or individuals to attain a political objective. Terrorism has been used by political organizations with both rightist and leftist objectives, by nationalistic and ethnic groups, by revolutionaries, and by government armies and secret police.

The basic concept behind an RDD is to disperse radioactive material in a location that will cause human health and environmental health harm, generate denial of access to areas, create confusion, disrupt command and communications, produce psychological harm in a population, and ultimately generate large detrimental economic consequences. The material can be dispersed into the atmosphere energetically by destroying a source package with conventional explosives (the typical "dirty bomb"), or a liquid/powder could be dispersed using sprayers. In addition to dispersal into the atmosphere, sources would also be dispersed in water supplies or members of the food chain. Although not a dispersal, a source could be covertly emplaced with the intent of irradiating people. Nuclear weapons or improvised nuclear devices are not discussed herein as their effects to the environment are discussed in sections on "Nuclear Weapons Testing" and "Nuclear Material Accidents".

The common elements for the most typical use of radionuclides for terrorist activities are shown in Fig. 41.

Technical Considerations for Environmental Contamination A radiological dispersal device (RDD) may be defined as:

- ► Any device, including any weapon or equipment, other than a nuclear explosive device, specifically designed to employ radioactive material by disseminating it to cause destruction, damage, or injury by means of the radiation produced by the decay of such material.

Almost any radioactive material can be used to construct an RDD, including fission products, spent fuel from nuclear reactors, and relatively low-level materials, such as medical, industrial, and research waste. Weapons grade materials (i.e., highly enriched uranium or plutonium) are not needed (nor desired) although they could be used. An RDD is designed to scatter radioactive material over a wide area, thereby contaminating the area.

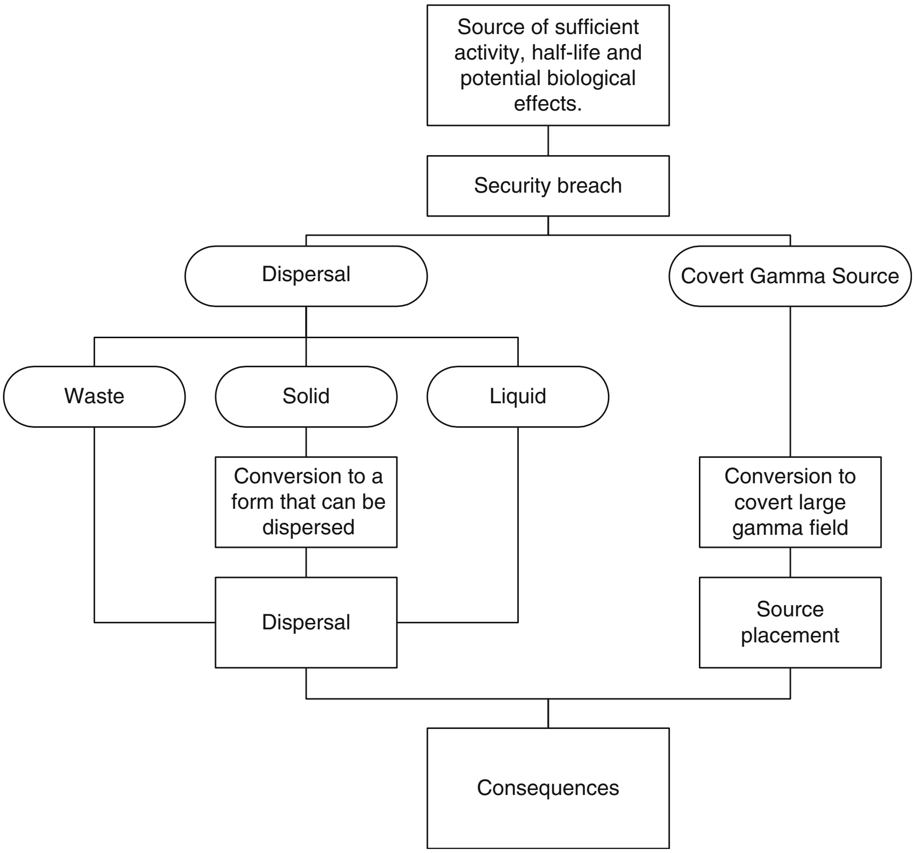
It is assumed that a quantity of radioactive material is in the possession of an individual or individuals for the purpose of constructing an RDD. That is, an individual is able (through some process) to obtain and store a source of radioactive material. Next, it is assumed that in order for an RDD to be effective in doing harm to human health, it must have a relatively large activity within a relatively small exposed population. Based upon the above assumptions, the following source classifications are high risk for an RDD:

- Large industrial-use sources of radioactive material (such as nondestructive testing gamma sources, thickness indicators, moisture detectors, and tank level monitors)
- Gamma sources used in radiotherapy medical facilities
- Various large sources of radioactivity used in government and commercial applications

It is worthy to note, however, that small, low-activity sources that would be ineffective at producing human bodily harm or environmental damage are capable of generating significant psychological damage to affected populations, and therefore even small radioisotope sources may be considered a threat.

There are four broad classifications with respect to exposure pathway:

1. Passive irradiation
2. Dissemination into a food or water chain directly


Source of sufficient activity, half-life and potential biological effects.
Security breach
Dispersal
Covert Gamma Source
Waste
Solid
Liquid
Conversion to a form that can be dispersed
Dispersal
Conversion to covert large gamma field
Source placement
Consequences

Radiation in the Environment, Sources of. Figure 41
Flowchart for possible terrorist activity involving radioactive material

3. Explosive aerosol dispersal
4. Nonexplosive aerosol dispersal

Passive irradiation involves placing a source in covert manner, as irradiating individuals or a population. This scenario affects a relatively small population and is readily detectable. The health physics theory behind external irradiation is well developed and readily applied to postulated scenarios. Dissemination into a food or water chain poses an ingestion threat to potentially large populations. Valued ecosystem component (VEC) analysis of radiological species is a well-developed science, primarily through pathways analysis as used in nuclear power plant operations. Explosive aerosol dispersal of radionuclide sources is a relatively undeveloped field of study. There are many parameters requiring investigation for explosive dispersal, including source metallurgical form, source encapsulation and explosive configuration. Nonexplosive dispersal of radiological contaminants is a relatively well-developed field of study, from the perspective of nuclear reactor accident plume characterization.

A large number of common radioisotopes have potential for generating either a radiological or psychological threat. Some of these are presented in Table 38. Although any radioisotope could be used in an RDD, the usefulness is usually determined by the availability, total activity, and radiological half-life.

A radiological dispersal device has never been used in war or terrorist act, and as such there is no operational data as to the effects on the environment. It would be expected that the effects would be similar to those from analogous events, such as nuclear material accidents (discussed in section on "Nuclear Material Accidents"), although targeted for optimum harm.

Radiation in the Environment, Sources of. Table 38
Potential industrial radioisotopes for RDD use

Radioisotope	Primary emission considered (secondary)	Typical use
H-3	β	Gaseous luminous light sources
Am-241	α (γ)	Smoke detectors
Cs-137	γ (β)	Irradiators
Co-60	γ (β)	Irradiators
Ir-192	γ (β)	Industrial radiography
Pu-238	α	Radioisotopic thermoelectric generators (RTG)
Ra-226	α (β, γ through progeny)	Historical medical and light sources
Sr-90	β	Radioisotopic thermoelectric generators (RTG)
Cf-252	Neutron (α, γ)	Oil well logging
Am-241/Be	Neutron (α, γ)	Oil well logging
Pu-238/Be	Neutron (α, γ)	Oil well logging

Future Directions

Environmental science is a constantly growing and adapting field. Although radionuclides have been in the environment since the origins of the universe, human understanding and impact assessment of environmental radionuclides has only taken place for approximately the past 100 years. Protection of the environment and stewardship over planetary resources is of great global concern. Some areas in which there is current and future need for research and development are listed below.

Dose–response models – Radionuclide environmental assessment ultimately relies upon risk estimation. Risk estimation has been, and continues to be, driven by the linear non-threshold (LNT) model. Despite the fact that it is widely acknowledged that LNT model has little data to support it below about 100 mGy dose, it continues to be used and drives all aspects of radiation protection from environmental restoration criteria, to human dosimetry, to the fundamental interactions with naturally occurring radioactive material. There is much research to be done to generate bona fide risk estimates from exposure at low dose (generally, the regime of environmental radioactivity).

Nuclear weapon reduction – Since the inception of nuclear weapons in 1945, there was a strong buildup and leveling off. Work is still being performed to completely ratify the Comprehensive Test Ban Treaty, with numerous political hurdles to overcome. Nuclear weapons have been the major contributor to worldwide distribution of anthropogenic radionuclides to the environment, so any efforts to eliminate testing are crucial to eliminate future introduction of anthropogenic radionuclides.

Radon dosimetry and mitigation – Radon is the major natural contributor to background dose, and, as such, efforts to measure radon both spatially and temporally will aid in radon mitigation strategies. Understanding radon dosimetry and the relationship with lung cancer and smoking is needed to determine if any detriment exists (observed, not calculated) from radon at low dose exposure.

TENORM – The issue of technologically enhanced naturally occurring radioactive material is both technical and regulatory in nature. There is much research to be done to reduce TENORM from chemical and industrial processes, as well as determining action and remediation levels in a reasonable and cost-effective manner.

Radioactive waste management – This has been one of the most studied areas related to environmental radioactivity for the past 30 years, and will continue to be a significant research area. Fundamental research is required in storage systems and strategies, modeling the fate of radionuclide transport far into the future, and dealing with issues such as transmutation of waste, waste reprocessing and recycling, and public acceptance issues. Other esoteric areas, such as revisiting the concepts of disposal in space or deep ocean disposal, require further exploration.

Nonhuman biota – Radiation protection principles have always worked on the belief that safe levels of radiation for humans did de facto protect the environment. More research is required examining explicit pathways, radionuclides, and environmentally valued ecosystem components to determine if this is indeed the case. Special attention must be given to vulnerable ecosystem populations.

Human and societal effect – Further research is required for understanding low dose radiation effects on populations, and exploring the paradoxes of high background radiation area populations and low cancer incidence. The health of the population (and nonhuman ecosystem components) requires complete epidemiological quantification, taking into account all endpoint confounders.

Contaminated ecosystem remediation – Decontamination and remediation of contaminated urban and rural environments is not a well-developed field, and there are a number of issues surrounding remediation efficacy, collection of remediated material, resuspension of contamination, and restoration guidelines all requiring further research. Immediate work is needed into assessing and quantifying sites of large introductions of historical anthropogenic radionuclides, such as the Arctic Ocean environment.

Modeling of the environment – Increases in computing power are related to increased ability to model complex systems such as the atmosphere and marine and terrestrial environments. There is much research to be done to model these complex systems. High-performance computing will enable complex compartmental systems to be modeled, which will enable prediction of environmental radioactivity transport over large spatial and temporal scales.

Mitigation of ground-level cosmogenic effects – There are numerous documented reports of solar storm activity disrupting ground-level electrical systems [108]. One major solar storm in 1859 allowed telegraph operators to continue utilizing their equipment despite unhooking them from battery power sources. Major solar storm activity is predicted for the 2012–2013 time frame, which in the worst case could disable numerous satellites, disrupt ground based communications, and trip electrical power distribution systems. The implications of losing electrical power, Internet, and audio/visual communications on a large public and governmental scale is obvious in a technology-driven society, and therefore strategies must be continuously developed for circumvention and mitigation of the effects in light of increasing advances in communication and electronics technology.

Orphan sources – Although largely an administrative issue, there are areas of research involving intelligent systems, data fusion, and other data analysis techniques that may assist in identifying probabilities of locating orphan sources, and for protecting against lost or orphan sources in the future.

Risk assessment - On 11 March 2011 an earthquake of magnitude 9 on the Richter scale struck off the east coast of Japan, proximal to Sendai. The earthquake, and subsequent tsunami, severely damaged the reactor buildings and control infrastructure for boiling water reactor (BWR) Units 1–6 at the Fukushima Daiichi nuclear reactor power facility. At the time of the event, Units 1–3 were operational and the earthquake triggered automatic shutdown of these units. Units 4–6 were off-line for inspection at the time of the incident, and unit 4 was completely defueled. Various hydrogen explosions occurred within the reactors' external structures and lack of water cooling caused damage to the reactors' fuel. Workers on-site were exposed to radioactivity primarily in contaminated water, and various isotopes were measured in and around environmental samples taken near the Fukushima Daiichi plant site [109]. The accident has been rated as Level 7 on the INES scale, which is the same level as was assigned to the Chernobyl NPP-4 accident. This "beyond-design-basis" accident is prompting regulators to reevaluate risk assessment models. Future work in redefining the design basis schema and details of consequence analysis from natural catastrophic events will be required for comprehensive protection of the environment as greater utilization of nuclear power is established.

In summary, radiation has provided much benefit to mankind over the past 100 years. Strong global stewardship and support is required to ensure that the supply of radionuclides, and the benefits of nuclear technology, remains safe into the future.

Bibliography

Primary Literature

1. NCRP (2009) Ionizing radiation exposure of the population of the United States. National Council on Radiation Protection and Measurements Report No. 160, Bethesda
2. Kathren R (1984) Radioactivity in the environment: sources, distribution and surveillance. Harwood Academic, London
3. Kendall G (2005) Factors affecting cosmic ray exposures in civil aviation. Int Congr Ser 1276:129–132
4. Bartlett D (2004) Radiation protection aspects of the cosmic radiation exposure of aircraft crew. Radiat Prot Dosim 109(4):349–355

5. Goldhagen P (2000) Overview of aircraft radiation exposure and recent ER-2 measurements. Health Phys 79(5): 526–544
6. Schraube H, Leuthold G, Roesler S, Heinrich W (1998) Neutron spectra at flight altitudes and their radiological estimation. Adv Space Res 21(12):1727–1738
7. Valković V (2000) Radioactivity in the environment. Elsevier Science, Amsterdam
8. Cooper J, Randall K, Sokhi R (2003) Radioactive releases in the environment: impact and assessment. Wiley, Etobicoke
9. Coogan L, Cullen J (2009) Did natural reactors form as a consequence of the emergence of oxygenic photosynthesis during the Achean? GSA Today 19(10):4–10
10. Meshik A, Hohenberg C, Pravdivtsea O (2004) Record of cycling operation of the natural nuclear reactor in the Oklo/Okelobondo area in Gabon. Phys Rev Lett 93(18):182302-1–182302-4
11. Gauthier-Lafaye F, Holliger P, Blanc P (1996) Natural fission reactors in the Franceville Basin, Gabon: a review of the conditions and results of a “critical event” in a geological system. Geochim Cosmochim Acta 60(23):4831–4852
12. Baum E, Knox H, Miller T (2002) Nuclides and isotopes, 16th edn. Lockheed Martin-Knolls Atomic Power Laboratory, Schenectady
13. WHO (1996) Indoor air quality: a risk-based approach to health criteria for radon indoors. WHO, Copenhagen
14. ICRP (1994) Protection against Radon-222 at home and at work. International Commission on Radiological Protection. ICRP Publication 65
15. Ghiassi-Nejad M, Beitolahi M, Fallahian N, Saghirzadeh M (2005) New findings in the very high natural radiation area of Ramsar, Iran. Int Congr Ser 1276:13–16
16. Magill J, Galy J (2005) Radioactivity radionuclides radiation. Springer, New York
17. ANSI (2009) Control and release of technologically enhanced naturally occurring radioactive material (TENORM) ANSI/HPS N13. American National Standards Institute
18. ICRP (1990) Recommendations of the International Commission on Radiological Protection. ICRP Publication 60. International Commission on Radiological Protection, Pergamon, Oxford
19. Goldstein B, Eisenbud M, Gesell T, Ibrahim S, Kocher D, Landa E, Pashoa A (1999) Evaluation of guidelines for exposures to technologically enhanced naturally occurring radioactive materials. National Academic, Washington, DC
20. Karam A, Vetter B (2009) Naturally occurring radioactive materials (NORM) and technologically enhanced NORM (TENORM). Medical Physics, Madison
21. Khater A (2004) Polonium-210 budget in cigarettes. J Environ Radioactiv 71:33–41
22. Skwarzec B, Ulatowski J, Struminska D, Borylo A (2001) Inhalation of Po-210 and Pb-210 from cigarette smoking in Poland. J Environ Radioactiv 57:221–230
23. Evans G (1993) Cigarette Smoking = Radiation hazards. Pediatrics 92(3):464–465
24. Abd El-Aziz N, Khater AEM, Al-Sewaidan HA (2005) Natural radioactivity contents in tobacco. Int Congr Ser 1276: 407–408
25. Eisenbud M, Gesell T (1997) Environmental radioactivity – from natural, industrial and military sources, 4th edn. Academic, Toronto
26. Landa E (1984) Geochemical and radiological characterization of soils from former radium processing sites. Health Phys 46(2):385–394
27. McBride J, Moore R, Whiterspoon J, Blanco R (1978) Radiological impact of airborne effluents of coal and nuclear plants. Science 202(4372):1045–1050
28. UNSCEAR (1988) Sources, effects and risks of ionizing radiation. United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR), 1988 Report to the General Assembly, with scientific annexes, United Nations, New York
29. Hedvall R (1996) Radioactivity concentrations in non-nuclear industries. J Environ Radioactiv 32(1–2):19–31
30. UNSCEAR (1993) Sources and effects of ionizing radiation. United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR), 1993 Report to the General Assembly, with scientific annexes, United Nations, New York
31. UNSCEAR (2000) Sources and effects of ionizing radiation, vol I and II. United Nations, New York
32. IAEA (2003) Radioactive waste management glossary. International Atomic Energy Agency. STI/PUB/1155, Vienna
33. IAEA (2009) Classification of radioactive waste. International Atomic Energy Agency. General Safety Guide GSG-1, Vienna
34. Salkeld R, Beichel R (1980) Nuclear waste disposal in space: implications of advanced space transportation. Acta Astronaut 7:1373–1387
35. Yablokov A (2001) Radioactive waste disposal is seas adjacent to the territory of the russian federation. Mar Pollut Bull 43(1–6):8–18
36. Champ M, Brooks J, Gomez L, Palmer H, Makeyev V, Betz F (2001) Ocean storage of nuclear waste? Experiences from the Russian Arctic. Mar Pollut Bull 43(1–6):1–7
37. Baxter M, Harms I, Osvath I, Povinec P, Scott E (1998) Modelling the potential radiological consequences of radioactive waste dumping in the Kara Sea. J Environ Radioactiv 39(2):161–181
38. Widner E, Flack S (2010) Characterization of the world’s first nuclear explosion, the trinity test, as a source of public radiation exposure. Health Phys 98(3):480–497
39. Parekh P, Semkow T, Torres M, Haines D, Cooper J, Rosenberg P, Kitto M (2006) Radioactivity in trinitite six decades later. J Environ Radioactiv 85:103–120
40. Glasstone S, Dolan P (1977) The effects of nuclear weapons, 3rd edn. US DoD, Washington, DC
41. Geddes D (1945) The atomic age opens. Pocket Books, New York
42. Loeber C (2002) Building the bombs: a history of the nuclear weapons complex. Sandia National Laboratories, Albuquerque
43. Salbu B (2008) Nuclear risks in Central Asia. In: Radioactive particles released from different nuclear sources: with focus

on nuclear weapons tests. Springer Science and Business Media, New York, Chapter 2

44. AFSWP (1951) Radiological defense, vol. II – the principles of military defense against atomic weapons. Armed Forces Special Weapons Project
45. Messenger G, Ash M (1986) The effects of radiation on electronic systems. Van Nostrand Reinhold, New York
46. CTBT Data obtained via Comprehensive Test Ban Treaty Organization. http://www.ctbto.org
47. Berger M, Jacobson D, Pinca S, Richards Z, Hess D, Hariss F, Page C, Peterson E, Baker N (2008) The state of coral reef ecosystems of the Marshall Islands. NOAA/NCCOS Center for Coastal Monitoring and Assessment's Biogeography Team, Silver Spring
48. Simon SL, Graham JC (1997) Findings of the first comprehensive radiological monitoring program of the Republic of the Marshall Islands. Health Phys 73(1):66–85
49. Cronkite EP, Conard RA, Bond VP (1997) Historical events associated with fallout from Bravo Shot–Operation Castle and 25 Y of medical findings. Health Phys 73(1):176–186
50. Richards ZT, Beger M, Pinca S, Wallace CC (2008) Bikini Atoll coral biodiversity resilience five decades after nuclear testing. Mar Pollut Bull 56(3):503–515
51. Robison WL, Conrado CL, Bogen KT, Stocker AC (2003) The effective and environmental half-life of 137Cs at Coral Islands at the former US nuclear test site. J Environ Radioactiv 69(3):207–223
52. Noshkin VE, Robison WL, Wong KM, Brunk JL, Eagle RJ, Jones HE (1997) Past and present levels of some radionuclides in fish from Bikini and Enewetak atolls. Health Phys 73(1):49–65
53. NCI (1997) Estimated exposures and thyroid doses received by the American people from Iodine-131 in fallout following Nevada atmospheric nuclear bomb tests. National Cancer Institute, Bethesda
54. Bowen S, Finnegan D, Thompson J, Miller C, Baca P, Olivas L, Geoffrion C, Smith D, Goishi W, Esser B, Meadows J, Namboodiri N, Wild J (2001) Nevada test site radionuclide inventory, 1951-1992. Los Alamos National Laboratory, LA-13859-MS, Los Alamos
55. Miller C, Bouville A (2005) Report on the feasibility of a study of the health consequences to the American population from nuclear weapons tests conducted by the Unites States and other nations. Centers for Disease Control, Atlanta
56. Frohmberg E, Goble R, Sanchez V, Quigley D (2000) The assessment of radiation exposures in native American communities from nuclear weapons testing in Nevada. Risk Anal 20(1):101–111
57. Russ A, George P, Goble R, Crema S, Liu C, Sachez D (2005) Native American exposure to 131I from nuclear weapons testing in Nevada. Hum Ecol Risk Assess 11(5):1047–1063
58. Haywood S, Smith J (1992) Assessment of potential doses at the Maralinga and emu test sites. Health Phys 63(6):624–630
59. Cooper M, Burns P, Tracy B, Wilks M, Williams G (1994) Characterization of plutonium contamination at the former nuclear weapons testing range at Maralinga in South Australia. J Radioanal Nucl Chem 177(1):161–184
60. Weber J, Till J (1999) Final report – Task 1: cleanup levels at other sites, radionuclide soil action level oversight panel. Risk Assessment Corporation, Neeses
61. Woollett S (2003) Rehabilitating Maralinga. Clean Air Environ Qual 37(4):31–33
62. Carter M, Moghissi A (1977) Three decades of nuclear testing. Health Phys 33:55–71
63. De Planque G (1998) The Mururoa study. IAEA Bull 40(4):21–23
64. Bourlat Y, Martin G (1992) Precise determination of the concentration of radiocesium in the water of Mururoa Lagoon. J Environ Radioactiv 17:13–29
65. Bourlat Y, Millies-Lacroix J, Nazard R (1995) Determination of plutonium radioactivity in Mururoa Lagoon water. J Radioanal Nucl Chem 192(2):387–408
66. Linsley G, McEwan A (1998) Potential doses at the atolls. IAEA Bull 40(4):38–42
67. Danesi P, Moreno J, Makarewicz M, Radecki Z (2002) Residual radioactivity in the terrestrial environment of the Mururoa and Fangataufa atolls nuclear weapons test sites. J Radioanal Nucl Chem 253(1):53–65
68. Egorov N, Novikov V, Parker F, Popov V (2000) The radiation legacy of the Soviet nuclear complex. Earthscan, London
69. Gusev B, Abylkassimova Z, Apsalikov K (1977) The Semipalatinsk NUCLEAR TEST SITE: a first assessment of the radiological situation and the test-related radiation doses in the surrounding territories. Radiat Environ Biophys 36:201–204
70. Grosche B (2002) Semipalatinsk test site: introduction. Radiat Environ Biophys 41:53–55
71. Carlsen T, Petersen L, Ulsh B, Werner C, Purvis K, Sharber A (2001) Radionuclide contamination at Kazakhstan's Semipalatinsk test site: implications on human and ecological health. Hum Ecol Risk Assess 7(4):943–955
72. Sakaguchi A, Yamamoto M, Hoshi M, Imanaka T, Apsalikov K, Gusev B (2006) Radiological situation in the vicinity of Semipalatinsk nuclear test site: Dolon, Mostik, Cheremushku and Budene settlements. J Radiat Res 47(Suppl A):A101–A116
73. Simon S, Baverstock K, Lindholm C (2003) A summary of evidence on radiation exposures near to the Semipalatinsk nuclear weapons test site in Kazakhstan. Health Phys 84(6):718–725
74. Balmukhanov S (2006) Medical effects and dosimetric data from nuclear tests at the Semipalatinsk test site. Defense Threat Reduction Agency. DRTA-TR-06-23, Fort Belvoir
75. Kemeny J (1979) Report of the President's commission on the accident at three mile Island – the need for change: the legacy of TMI., Library of Congress Catalog Card No. 79-25694. ISBN 0-935758-00-3
76. Waller E, Cole D (1999) An environmental radionuclide baseline study (ERBS) near three Canadian Naval Ports. Health Phys 77:37–42
77. Strålskyddsnytt (2006) Tjernobyl 20 år. Strålskyddsnytt 1, 11.
78. Borovoy A, Beskorovainyi V, Bogatov S, Vysotskiy E, Gavrilov S, Ivanov A, Kotovich V, Krinitsyn A, Molodykh V, Nemchinov Y,

Pazukhin E, Rudko V, Sharovarov G, Shcherbin V (1998) The shelter's current safety status and situation development forecasts (updated version). Tacis Services DG IA, Brussels

79. McLaughlin T, Monahan S, Pruvost N, Frolov V, Ryazanov B, Sviridov V (2000) A review of criticality accidents. Los Alamos National Laboratory, Los Alamos
80. IAEA (1998) The radiological accident in the reprocessing plant at Tomsk. International Atomic Energy Agency. STI/PUB/1060, Vienna
81. Usachev V, Markov G (2003) Incidents caused by red oil phenomena at semi-scale and industrial radiochemical units. Radiochemistry 45(1):1–8
82. Anspaugh L, Degteva M, Vasilenko E (2002) Mayak production association: introduction. Radiat Environ Biophys 41:19–22
83. IAEA (2009) INES: The international nuclear and radiological event scale user's manual, 2008 ed. IAEA, Vienna
84. Moran B (2009) The day we lost the H-Bomb: cold war, hot nukes and the worst nuclear weapons disaster in history. Ballantine Books, New York
85. Place W, Cobb F, Defferding C (1975) Palomares summary report. Field Command, Defense Nuclear Agency, Technology and Analysis Directorate, Kirtland Air Force Base, New Mexico
86. Jiménez-Ramos M, García-Tenorio R, Vioque I, Manjón G, García-León M (2006) Presence of plutonium contamination in soils from Palomares (Spain). J Environ Radioactiv 142:487–492
87. Espinosa A, Aragon A, Stradling N, Hodgson A, Birchall A (1998) Assessment of doses to adult members of the public in Palomares from inhalation of plutonium and Americium. Radiat Prot Dosim 79(1–4):161–164
88. Erisken V (1990) Sunken nuclear submarines – a threat to the environment? Norwegian University Press, Oslo
89. Waller E, Cole D, Jamieson T (2008) Radiation protection issues related to Canadian museum operations. Health Phys 94(2 Suppl 1):S155–S159
90. NCRP (2007) Development of a biokinetic model for radionuclide-contaminated wounds and procedures for their assessment, dosimetry and treatment. National Council on Radiation Protection and Measurements. Report No. 156, Bethesda
91. Burkart W, Danesi P, Hendry J (2005) Properties, use and health effects of depleted uranium. Int Congr Ser 1276:133–136
92. Papastefanou C (2002) Depleted uranium in military conflicts and the impact on the environment. Health Phys 83(2):280–282
93. McLaughlin J (2005) Public health and environmental aspects of DU. Int Congr Ser 1276:137–140
94. Aftergood S (1989) Background on space nuclear power. Sci Glob Secur 1:93–107
95. Bennett G (2006) Space nuclear power: opening the final frontier, American Institute of Aeronautics and Astronautics AIAA 2006-4191. In: Presentation at 4th international energy conversion engineering conference and exhibit (IECEC), San Diego, pp 12–13
96. Hardy E, Krey P, Volchok H (1973) Global inventory and distribution of fallout plutonium. Nature 241:244–245
97. Gummer W, Campbell F, Knight G, Richard J (1980) Cosmos 954 – the occurrence and nature of recovered debris. Atomic Energy Control Board. AECB- INFO-0006
98. IAEA (2000) Categorization of radiation sources. International Atomic Energy Agency. IAEA GOV/2000/34-GC(44)/7 Attachment 3, Vienna
99. NN (2003) RDD Report offers steps to reduce threat. *Nuclear News* 03/2003, 37
100. Ferguson C, Kazi T, Perera J (2003) Commercial radioactive sources: surveying the security risks. Occasional Paper No. 11, Centre for Non-Proliferation Studies, Monterey Institute of International Studies, Monterey
101. Saha G (1998) Fundamentals of nuclear pharmacy, 4th edn. Springer, New York
102. Fischer H, Ulbrich S, Pittauerová D, Hettwig B (2009) Medical radioisotopes in the environment – following the pathway from patient to river sediment. J Environ Radioactiv 100(12):1079–1085
103. NCRP (1987) Radiation exposure of the US population from consumer products and miscellaneous sources. National Council on Radiation Protection and Measurements. Report No. 95, Bethesda
104. Bevelacqua J (1995) Contemporary health physics: problems and solutions. Wiley-Interscience, Toronto
105. IAEA (2002*)* Inadequate control of world's radioactive sources. International Atomic Energy Agency. IAEA Press Release 2002/09, Vienna
106. IAEA (2004) Strengthening control over radioactive sources in authorized use and regaining control over orphan sources – national strategies. International Atomic Energy Agency. IAEA-TECDOC-1388, Vienna
107. IAEA (1988) The radiological accident in Goiania. International Atomic Energy Agency. STI/PUB/815, Vienna
108. Boteler D, Pirjola R, Nevanlinna H (1998) The effects of geomagnetic disturbances on electrical systems at the Earth's surface. Adv Space Res 22(1):17–27
109. Fukushima (2011) Special report: Fukushima Daiichi after the earthquake and tsunami. Nuclear News 54:17–18b, 83–84b

Books and Reviews

Aycik G, Ercan A (1997) Radioactivity measurements from coals and ashes from coal fired power plants from the southwestern part of Turkey. J Environ Radioactiv 35:23–35

Bailey R, Clark H, Ferris J, Krause S, Strong R (2002) Chemistry of the environment, 2nd edn. Academic, New York

Baryakhtar V, Kukhar V, Los I, Poyarkov V, Kholosha V, Shestopalov V (1998) Comprehensive risk assessment of the consequences of the chernobyl accident. Science and Technology Center in Ukraine – Ukrainian Radiation Training Centre. Project No. 369. Kiev, Ukraine

Brodsky A (1996) Review of radiation risks and uranium toxicity with application to decisions associated with decommissioning clean-up criteria. RSA, Hebron

Bussard RW, DeLauer RD (1958) Nuclear rocket propulsion. McGraw Hill, Toronto

Calomet D (1989) Ocean disposal of radioactive waste: status report. IAEA Bull 4:47–50

Cember H, Johnson T (2009) Introduction to health physics, 4th edn. McGraw Hill, Toronto

Chernobyl (2005) Chernobyl's legacy: health, environmental and socio-economic impacts and recommendations to the Governments of Belarus, the Russian Federation and Ukraine. The Chernobyl Forum: 2003-2005, 2nd rev edn

Crick MJ, Linsley GS (1982) An assessment of the radiological impact of the windscale reactor fire, October 1957. Report NRPB-R135, National Radiological Protection Board, Chilton

Dai L, Wei H, Wang L (2007) Spatial distribution and risk assessment of radionuclides in soils around a coal-fired power plant: a case study from the city of Baoji. China Environ Res 104: 201–208

Fry RJ (2000) Radiation protection guidance for activities in low-earth orbit. National Council on Radiation Protection and Measurements. NCRP Report No. 132, Bethesda

Gomez M (1981) Radiation hazards in mining: control, measurement, and medical aspects. Society of Mining Engineers. American Institute of Mining, Metallurgical and Petroleum Engineers, New York

Heaps L (1978) Operation morning light. Ballantine Books, Toronto

IAEA (1996) Issues in radioactive waste disposal – second report of the working group on principals and criteria for radioactive waste disposal. International Atomic Energy Agency. IAEA-TECDOC-909, Vienna

IAEA (1999) Protection of the environment from the effects of ionizing radiation – a report for discussion. International Atomic Energy Agency. IAEA-TECDOC-1091, Vienna

Journal of Environmental Radioactivity. Elsevier Science, Maryland Heights, MO. ISSN: 0265-931X

Lamarsh J, Baratta A (2001) Introduction to nuclear engineering, 3rd edn. Prentice Hall, Toronto

Lehr J, Hyman M, Gass T, Seevers W (2002) Handbook of complex environmental remediation problems. McGraw Hill, Toronto

Lewis WB, Ward AC (1953) An appreciation of the problem of reactor shut-off rods with special reference to the NRX reactor. Atomic Energy of Canada Limited. AECL No. 590, Chalk River

Louvar J, Louvar B (1998) Health and environmental risk analysis. Prentice Hall, Toronto

Mangeno J, Steele J, Betschart J (1991) Environmental monitoring and disposal of radioactive wastes from U.S. Naval nuclear-powered ships and their support facilities 1990. Naval Nuclear Propulsion Program, Department of the Navy. Report NT-91-1

Mangeno J, Tyron A, Burrows C (1991) Occupational radiation exposure from U.S. Naval nuclear propulsion plants and their support facilities. Naval Nuclear Propulsion Program, Department of the Navy. Report NT-91-2

May J (1989) The Greenpeace handbook of the nuclear age. McClelland and Stewart, Toronto

McCracken G, Stott P (2005) Fusion: the energy of the universe. Elsevier Academic, New York

Messenger GC, Ash MS (1997) Single event phenomena. Chapman & Hall, Toronto

Moeller D (2005) Environmental health, 3rd edn. Harvard University Press, Cambridge

Moeller D (2005) Environmental health physics: 50 years of progress. Health Phys 88:676–696

Moore T, Dietrich D (1987) Chernobyl and its legacy. EPRI J 5:21

Oberg J (1999) The probe that fell to Earth. *New Scientist* 6 March 1999

Paić G (1988) Ionizing radiation: protection and dosimetry. CRC, Boca Raton

Reese H (1986) DoD nuclear mishaps. Armed Forces Radiobiology Research Institute. AFRRI SP86-2, Bethesda

Rippon S (1986) Chernobyl: The Soviet report. Nucl News 1:8

Scott RL (1971) Fuel-melting incident at the fermi reactor on Oct 5, 1966. Nucl Saf 12:123–134

Silva P, Camargo I, Mazzilli B (2006) Radioactivity of coal and ashes from figueira coal power plant in Brazil. J Radioanal Nucl Chem 270:597–602

Simon SL (1997) A brief history of people and events related to atomic weapons testing in the Marshall Islands. Health Phys 73(1):5–20

Till J, Grogan H (2008) Radiological risk assessment and environmental analysis. Oxford University Press, Toronto

Turner J (1995) Atoms, radiation and radiation protection, 2nd edn. Wiley Interscience, Toronto

Tykva R, Berg D (eds) (2004) Man-made and natural radioactivity in environmental pollution and radiochronology. Kluwer, Dordecht/Holland

Zeevaert T, Sweeck L, Vanmarcke H (2006) The radiological impact from airborne routine discharges of a modern coal-fired power plant. J Environ Radioactiv 85:1–22

Radiation Shielding

J. Kenneth Shultis[1], Richard E. Faw[2]
[1]Department of Mechanical & Nuclear Engineering,
Kansas State University, Manhattan, KS, USA
[2]Department of Mechanical & Nuclear Engineering,
Kansas State University, Winston Salem, NC, USA

Article Outline

Glossary
Definition of the Subject
Introduction
History of Shielding
Practice of Radiation Shielding
Basic Analysis Methods

Intermediate Methods for Photon Shielding
Intermediate Methods for Neutron Shielding
Gamma-Ray and Neutron Reflection
Radiation Streaming Through Ducts
Gamma-Ray and Neutron Skyshine
Transport Theory
Future Directions
Bibliography

Glossary

Albedo A quantity describing how neutrons or photons incident on the surface of some medium (e.g., a wall) are reflected or reemitted from the surface.

Buildup factor A factor to account for production of secondary photons in a shield. The transmitted dose from only uncollided photons times the buildup factor equals the dose from all photons, uncollided plus secondary photons.

Dose A general term for the energy transferred from radiation to matter. Specifically, the absorbed dose is the amount of energy imparted to matter from ionizing radiation in a unit mass of that matter. Units are the gray (Gy) and rad, respectively, equivalent to 1 J/kg and 100 ergs/g.

Flux A measure of the intensity of a radiation field. Specifically, it equals the number of radiation particles entering, in a unit time, a sphere of cross-sectional area ΔA divided by ΔA, as $\Delta A \to 0$. The flux, integrated over a specified time interval is called the *fluence*.

Interaction coefficient A quantity, denoted by μ, describing how readily a photon or neutron interacts with a given medium. Specifically, it is the probability a radiation particle of energy E will interact in a specified manner per unit distance of travel, for infinitesimal distances. It thus has units of inverse length. The total interaction coefficient $\mu = \sum_i \mu_i$ where μ_i is the coefficient for the ith type of interaction (e.g., scattering, absorption).

Neutron A neutral subatomic particle that collectively with positively charged protons forms an atomic nucleus. Although both are composite particles composed of quarks and gluons, for the energies considered in this entry they can be viewed as fundamental unchangeable particles.

Photon A quantum of electromagnetic radiation with energy $E = hv$, where h is Planck's constant and v is the frequency. Photons produced by a change in the structure of the nucleus are called *gamma photons* and those produced by atomic electron rearrangement are called *x-rays*.

Skyshine A term for the radiation that reaches some point of interest after being scattered by the atoms in the atmosphere back to the point of interest.

Transport equation Also known as the linearized Boltzmann equation, it describes rigorously the spatial, energy, and angular distribution of neutrons or photons in any medium with arbitrary source distributions. From its solution, the radiation flux or dose anywhere in the medium can be determined.

Definition of the Subject

We live in a world that abounds in radiation of all types. Many radiations, such as the neutrinos or visible light from our sun present little risk to us. Other radiations, such as medical x-rays or gamma rays emitted by radioactive materials, have the potential to cause us harm. In this entry, only the transport of *indirectly ionizing radiation* is considered. These radiations consist of chargeless particles such as neutrons or photons that, upon interacting with matter, produce energetic secondary charged particles called *directly ionizing radiation*. It is these secondary charged particles that, through ionization and excitation of ambient atoms along their paths, cause radiation damage to biological tissues or other sensitive materials.

To mitigate radiation damage, a *shield* is often interposed between a source of ionizing radiation and the object to be protected so that the radiation levels near the object are reduced to tolerable levels. Typically, a shield is composed of matter that effectively diminishes the radiation that is transmitted. (However, there are noncorporeal shields such as magnetic fields that deflect moving charged particles. The earth's magnetic field serves as such a shield to protect us from charged particles reaching earth from outer space.) The term *radiation shielding* refers usually to a system of shields constructed for a specific radiation protection purpose. The term also refers to the study of shields – the topic of this entry.

Introduction

The origins of shielding go back to the science of optics in which the exponential attenuation of light was long recognized. The exponential attenuation of radiation rays is still widely used for neutron and photon shielding. Also the governing field equation that describes how radiation migrates through matter was introduced in 1872 by Ludwig Boltzmann who used it to study the kinetic theory of gas. All this occurred before the discovery of ionizing radiation! The *radiation transport equation* is just a special case of the Boltzmann equation applied to situations in which radiation particles do not interact among themselves.

The study of shielding has many aspects: transport of (deeply penetrating) indirectly ionizing radiation in the shield, the production of very slightly penetrating secondary (directly ionizing) radiation in the shield and its surroundings, the radiation levels in the vicinity of the shield, deposition of heat in the shield, radiation penetration through holes in the shield, radiation scattered around the shield, selection of shielding materials, optimization of the shielding configuration, and the economics of shield design. It also involves understanding of related matters such as radiation source characteristics, radiation protection standards, and the fundamentals of how radiation interacts with matter.

The restriction of this entry to indirectly ionizing radiation is of a practical nature. Sources of charged particles, such as the alpha and beta particles emitted in some types of radioactive decay, can and do cause biological damage, particularly if the radioactive material is ingested. Here, however, it is assumed that the radiation sources are external to the body or the sensitive material of interest. Such external sources also usually emit far more penetrating indirectly ionizing radiation, and any shield that is effective against indirectly ionizing radiation is usually more than adequate to stop the directly ionizing radiation.

History of Shielding

To appreciate better the current state of shielding practice, it is important to understand how the discipline developed and what were the driving forces that caused it to mature. In this section, a brief overview of the history of shielding is presented. (A greatly expanded version of the following synopsis is provided by Shultis and Faw [1].)

Early History

The hazards of x-rays were recognized within months of Roentgen's 1895 discovery, but dose limitation by time, distance, and shielding was at the discretion of the individual researcher until about 1913. Only then were there organized efforts to create groups to establish guidelines for radiation protection. And it was not until about 1925 that instruments became available to quantify radiation exposure.

In 1925, Mutscheller [2] introduced important concepts in x-ray shielding. He expressed the erythema dose (An *ED* value of unity represents a combination of time, distance, and beam current just leading to a first-degree burn) *ED* quantitatively in terms of the beam current i (mA), exposure time t (min), and source-to-receiver distance r (m), namely, $ED = 0.00368it/r^2$, independent of x-ray energy. Mutscheller also published attenuation factors in lead as a function of lead thickness and x-ray average wavelength.

Evolutionary changes to x-ray shielding were made during the decades preceding World War II. These included consideration of scattered x-rays, refinements in shielding requirements in terms of x-ray tube voltages, recommendations for use of goggles (0.25-mm Pb equivalent) and aprons (0.5-mm Pb equivalent) for fluoroscopy, and specifications for tube-enclosure shielding and structural shielding for control rooms.

The other major source of ionizing radiation before World War II was the medical and industrial use of radioactive radium discovered by Marie and Pierre Curie in 1898. Not until 1927 were lead shielding standards recommended for radium applicators, solutions, and storage containers. For example, the International X-Ray and Radium Protection Committee recommended that tubes and applicators should have at least 5 cm of lead shielding per 100 mg of radium. It was not until 1941 that a tolerance dose for radium, expressed in terms of a maximum permissible body burden of 0.1 μCi, was established. This was done largely in consideration of the experiences of early "radium-dial" painters and the need for standards on safe handling of radioactive luminous compounds [3].

Manhattan Project and the Early Postwar Period

Early reactor shielding. During World War II, research on nuclear fission, construction of nuclear reactors, production of enriched uranium, generation of plutonium and its separation from fission products, and the design, construction, testing, and deployment of nuclear weapons all were accomplished at breakneck speed in the Manhattan Project. Radiation sources new in type and magnitude demanded not only protective measures such as shielding but also examination of biological effects and establishment of work rules.

The construction of nuclear reactors for research and for plutonium production required shield designs for both gamma rays and neutrons. However, with only sparse empirical data and large uncertainties about how neutrons and gamma rays migrate through shields, shield designers acted very conservatively. For example, shielding for both Fermi's 1943 graphite pile in Chicago and the 1947 X-10 research reactor at what is now Oak Ridge National Laboratory was adequate for gamma rays and overdesigned for neutrons. Operation of the X-10 reactor, built to provide data for the design of plutonium-production reactors, revealed problems with streaming of gamma rays and neutrons around access holes in the shield. The water-cooled graphite plutonium-production reactors at Hanford, Washington used iron thermal shields and high-density limonite and magnetite concrete as biological shields.

By the 1940s, the importance of scattered gamma rays was certainly known from measurements, and use of the term *buildup factor* to characterize the relative importance of scattered and unscattered gamma rays had its origin during the days of the Manhattan Project. Neutron diffusion theory and Fermi age theory were established, but shielding requirements for high-energy neutrons were not well understood. Wartime radiation shielding was an empirical, rule-of-thumb craft.

Nuclear reactors for propulsion. The Atomic Energy Act of 1946 transferred control of nuclear matters from the Army to the civilian Atomic Energy Commission (AEC). That same year, working with the AEC, the US Navy began development of a nuclear powered submarine and the US Air Force, a nuclear powered aircraft. Both of these enterprises demanded minimization of space and weight of the nuclear-reactor power source. Such could be accomplished only by minimizing design margins and that required knowledge of mechanical, thermal, and nuclear properties of materials with greater precision than known before.

Research reactors were constructed at various national laboratories in the USA and Britain to provide the much needed shielding data. The first such research program was begun in 1947 at Oak Ridge National Laboratory with the construction of the X-10 graphite reactor. The X-10 graphite reactor had a 2-ft square aperture in its shielding from which a neutron beam could be extracted, the intensity being augmented by placement of fuel slugs in front of the aperture. Attenuation of neutrons could then be measured within layers of shielding materials placed against the beam aperture. Early measurements revealed the importance of capture gamma rays produced when neutrons were absorbed. Improved experimental geometry was obtained by using a converter plate containing enriched uranium instead of relying on fission neutrons from fuel slugs. A broadly uniform beam of thermal neutrons incident on the plate generated a well-defined source of fission neutrons. A water tank was adjacent to the fission source, with shielding slabs and instrumentation within the tank. This *Lid Tank Shielding Facility*, LTSF, was the precursor of many so-called *bulk-shielding facilities* incorporated into many water-cooled research reactors.

Although a nuclear powered aircraft never flew, the wealth of information gained on the thermal, mechanical, and shielding properties of many special materials is a valuable legacy. To obtain shielding data in the absence of ground reflection of radiation, several specialized facilities were constructed. A test reactor was suspended by crane for tests of ground reflection. Then an aircraft shield test reactor was flown in the bomb-bay of a B-36 aircraft to allow measurements at altitude. The Oak Ridge tower shielding facility (TSF) went into operation in 1954, and remained in operation for almost 40 years. Designed for the aircraft nuclear propulsion program, the facility allowed suspension of a reactor hundreds of feet above grade and separate suspension of aircraft crew compartments. In its long life, the TSF also supported nuclear defense and space nuclear applications.

Streaming of radiation through shield penetrations and heating in concrete shields due to neutron and gamma-ray absorption were early shielding studies conducted in support of gas-cooled reactor design.

Additional efforts were undertaken soon thereafter at universities as well as government and industrial laboratories. Shielding material properties, neutron attenuation, the creation of capture and inelastic scattering gamma rays, reflection and streaming of neutrons and gamma rays through ducts and passages, and radiation effects on materials were major research topics.

The Decade of the 1950s

This era saw the passage in the USA of the Atomic Energy Act of 1954, the Atoms for Peace program, and the declassification of nuclear data. During this decade, many simplified shielding methods were developed that were suitable for hand calculations. The first digital computers appeared and were quickly used for radiation transport calculations. The US Air Force also started a short-lived nuclear rocket program.

Advances in neutron shielding methods. These advances resulted from measurements at the LTSF and other bulk-shielding facilities. One advancement was the measurement of point kernels, or Green's functions, for attenuation of fission neutrons in water and other hydrogenous media. The other was the discovery that the effect of water-bound oxygen, indeed the effect of homogeneous or heterogeneous shielding materials in hydrogenous media, could be modeled by exponential attenuation governed by effective "removal" cross sections for the non-hydrogen components. The LTSF allowed measurement of removal cross section for many materials.

Advances in gamma-ray shielding methods. As the decade began, researchers at the National Bureau of Standards investigated electron and photon transport. Much of this effort dealt with the moments method for solving the transport equation describing the spatial, energy, and angular distributions of radiation particles emitted from fixed sources. From such calculations, *buildup factors* to account for scattered photons were determined for various shielding media and shield thicknesses. Various empirical formulas were also developed to aid in the interpolation of the buildup-factor data.

Advances in Monte Carlo computational methods. The Monte Carlo method of simulating radiation transport computationally has its roots in the work of John von Neumann and Stanislaw Ulam at Los Alamos in the 1940s. Neutron-transport calculations were performed in 1948 using the ENIAC digital computer which had commenced operations in 1945. In this decade, major theoretical advances in Monte Carlo methods were made and many clever algorithms were invented to allow Monte Carlo simulations of radiation transport through matter. Little did the pioneers of this transport approach realize that Monte Carlo techniques would become indispensable in modern shielding practice.

The Decade of the 1960s

The 1960s saw the technology of nuclear-reactor shielding consolidated in several important publications. Blizard and Abbott [4] edited and released a revision of a portion of the 1955 Reactor Handbook as a separate volume on radiation shielding, recognizing that reactor shielding had emerged from nuclear-reactor physics into a discipline of its own. In a similar vein, the first volume of the Engineering Compendium on Radiation Shielding [5] was published. These two volumes brought together contributions from scores of authors and had a great influence on both practice and education in the field of radiation shielding.

This exciting decade also saw the beginning of the Apollo program, the start of the NASA NERVA (Nuclear Engine for Rocket Vehicle Application) program, the deployments in space of SNAP-3, a radioisotope thermoelectric generator in 1961, and SNAP-10A nuclear-reactor power system in 1965. It also saw the Cuban missile crisis in October 1962 and a major increase in the cold-war apprehension about possible use of nuclear weapons. The Apollo program demanded attention to solar flare and cosmic radiation sources and the shielding of space vehicles. Cold-war concerns demanded attention to nuclear-weapon effects, particularly structure shielding from nuclear-weapon fallout. Reflection of gamma rays and neutrons and their transmission through ducts and passages took on special importance in structure shielding. The rapid growth in access to digital computers allowed introduction of many computer codes for shielding design and fostered advances in solving various approximations to the Boltzmann transport equation for neutrons and gamma rays. Similar advances were made in treating the slowing down and transport of charged particles.

Space shielding. Data gathered over many years revealed a very complicated radiation environment in space. Two trapped-radiation belts had been found to surround the earth, an inner proton belt and an outer electron belt. Energy spectra and spatial distributions in these belts are determined by the earth's magnetic field and by the solar wind, a plasma of low-energy protons and electrons. The radiations pose a risk to astronauts and to sensitive electronic equipment. Uniform intensities of very-high-energy galactic cosmic rays demand charged-particle shielding for protection of astronauts in long duration missions. The greatest radiation risk faced by Apollo astronauts was from solar flare protons and alpha particles with energies as great as 100 MeV for the former and 400 MeV for the latter. The overall subject of space radiation shielding is treated by Haffner [6].

Structure shielding. Structure shielding from nuclear-weapon fallout required careful examination of the atmospheric transport of gamma rays of a wide range of energies and expression of angular distributions and related data in a manner easily adopted to analysis of structures. There was a need to assess, at points within a structure, the ratios of interior dose rates to that outside the building, called *reduction factors*. These factors were measured experimentally and also calculated with the transport moments method which had been used so successfully in calculation of buildup factors.

Other shielding advances. Of great importance to structure shielding, but also of interest in reactor and nuclear plant shielding, were the development of simplified methods to quantify neutron and gamma-ray streaming through ducts and voids in shields. This decade saw the development of removal-diffusion methods to describe quite accurately the penetration and slowing down of fast fission neutrons in shields. Finally, a simplified approach was developed to describe how gamma rays or neutrons incident on some material are scattered back. The central concept in this approach is the *particle albedo*, a function that describes how radiation incident on a thick medium, a concrete wall for example, is reemitted or reflected back from the surface. Measurements, theoretical calculations, and approximating formulas for both neutron and gamma-ray albedos were developed in this decade.

Digital computer applications. Radiation transport calculations are by nature very demanding of computer resources. The community of interest in radiation transport and shielding has been served magnificently for more than 4 decades by the Radiation Safety Information Computational Center (RSICC). Established in 1962 as the Radiation Shielding Information Center (RSIC) at Oak Ridge National Laboratory, RSICC's mission is to provide in-depth coverage of the radiation transport field to meet the needs of the international shielding community.

The 1960s saw many new "mainframe" computer codes developed and disseminated. Among these codes were gamma-ray "point-kernel" codes such as ISOSHLD and QAD, with versions of both still in use after almost 4 decades. The discrete-ordinates method of solving the Boltzmann transport equation was devised in the 1950s and put into practice in the 1960s in a series of computer codes, such as DTF, DOT, and ANISN. The spherical harmonics method of treating neutron spatial and energy distributions in shields was advanced by Shure [7] in one-dimensional P_3 calculations. Progress in Monte Carlo methods advanced in pace with discrete-ordinates methods, and the multigroup Monte Carlo code for neutron and gamma-ray transport, MORSE, was introduced at the end of the decade. The continuous energy Monte Carlo code, now known as MCNP, also began in this decade at Los Alamos National Laboratory. A general-purpose particle-transport code MCS was written in 1963 to be followed by the MCN code for three-dimensional calculations written in 1965.

The Decade of the 1970s

The Nuclear Non-Proliferation Treaty (NPT) of 1968 and the National Environmental Policy Act (NEPA) of 1969 had major impacts on the radiation shielding field in the 1970s and succeeding decades. The NPT precluded nuclear fuel reprocessing and led to ever-increasing needs for on-site storage of spent fuel at nuclear power plants. NEPA required exhaustive studies of off-site radiation doses around nuclear power plants and environmental impacts of plant operations. Early in the 1970s, there were major disruptions in oil supplies caused by the OPEC embargo. The response in the USA was an energy policy that forbade electricity production using oil or natural gas. The result was placement of many orders for nuclear power plants despite NPT and NEPA constraints. In the field of radiation shielding, special attention was given to plant design issues such as streaming of neutrons and gamma rays through voids, passageways, and shield penetrations, and to operational issues such as fission-product inventories in fuels and gamma-ray skyshine, particularly associated with ^{16}N sources.

Information essential for plant design, fuel management, and waste management is data tracking radionuclide activities in reactor fuel and process streams, and corresponding strengths and energy spectra of sources, including fission products, activation products, and actinides. To accomplish this, the ORIGEN codes were developed at Oak Ridge National Laboratory and the CINDER code was developed at Los Alamos National Laboratory. Assessment of radiation doses from airborne beta-particle emitters was also studied for the first time. Although the ETRAN Monte Carlo code for electron transport was available at the National Bureau of Standards, work began in the mid-1970s at Sandia Laboratory on the TIGER code and at Stanford Linear Accelerator Center on the EGS code, both for coupled photon and electron transport by the Monte Carlo method.

Design needs brought new attention to buildup factors and to attenuation of broad beams of neutrons and gamma rays. Definitive compilations were made of buildup factors and also the attenuation and reflection by shields obliquely illuminated by photons. Detailed results were also obtained for transmission of neutrons and secondary gamma rays through shielding barriers. This decade also saw the publication of two important NCRP reports [8,9] dealing with neutron shielding and dosimetry and with design of medical facilities that protected against effects of gamma rays and high-energy x-rays.

Design and analysis needs also fostered continuing attention to computer codes for criticality and neutron-transport calculations. A series of more robust discrete-ordinates transport codes were developed. Advances in Monte Carlo calculations were also made. The MCN code was merged with the MCG code in 1973 to form the MCNG code for treating coupled neutron–photon transport. Another merger took place with the MCP code in 1977, allowing detailed treatment of photon transport at energies as low as 1 keV. This new code was known, then and now, as MCNP.

The 1980s and 1990s

These years saw the consolidation of resources for design and analysis work. In the 1980s, personal computers allowed methods such as point-kernel calculations to be programmed. In the 1990s, personal computers took over from the mainframe computers in even the most demanding shielding design and analysis. Comprehensive sets of fluence-to-dose conversion factors became available for widespread use. Radionuclide decay data became available in databases easily used for characterizing sources. Gamma-ray buildup factors were computed with precision and a superb method of data fitting was devised. All these carried point-kernel as well as more advanced shielding methodology to a new plateau.

Databases. Kocher [10] published radioisotope decay data for shielding design and analysis that largely supplanted earlier compilations. Then a new MIRD compendium [11] and ICRP-38 database became the norms, with the latter especially useful for characterizing low-energy x-ray and Auger electron emission. Today a wealth of nuclear structure and decay data is available on the web from the National Nuclear Data Center at Brookhaven National Laboratory (http://www.nndc.bnl.gov/index.jsp).

Advances in buildup factors. Refinements in the computation of buildup factors continued to be made over the years. Computer codes now could account for not only Compton scattering and photoelectric absorption, but also positron creation and annihilation, fluorescence, and bremsstrahlung. Calculation of buildup factors incorporating all these sources of secondary photon radiation was made leading to a comprehensive set of precise buildup factors standardized for use in design and analysis [12]. Also a new five-parameter buildup-factor formulation, called the *geometric progression* formula, was introduced. Although difficult to use for hand calculations, it is an extraordinarily precise formula and is today used in most modern point-kernel codes. Both the calculated buildup factors and the coefficients for the geometric progression buildup factors are tabulated in the design standard [12].

Cross sections and dose conversion factors. Authoritative cross-section data are now available in the ENDF/B (evaluated nuclear data file) (http://www.nndc.bnl.gov/exfor/endf00.htm) database containing evaluated cross sections, spectra, angular distributions, fission product yields, photo-atomic and thermal scattering law data, with emphasis on neutron-induced reactions. The National Institute of Science and Technology (NIST) has long been the repository for gamma-ray interaction coefficients. The Institute also sponsors the XCOM cross-section code, which may be executed on the NIST Internet site (http://physics.nist.gov/PhysRefData/Xcom/Text/XCOM.html) or downloaded for personal use.

Gamma-ray fluence-to-dose conversion factors for local values of exposure or kerma may be computed directly from readily available energy transfer or energy absorption coefficients for air, tissue, etc. Neutron conversion factors for local values of tissue kerma were computed by Caswell et al. [13]. As the second century of radiation protection begins, there are two classes of fluence-to-dose conversion factors in use for neutrons and gamma rays. One very conservative class is to be used for operational purposes at doses well below regulatory limits. This class is based on doses at fixed depths in 30-cm diameter spherical phantoms irradiated in various ways. The other class is to be used for dose assessment purposes, and not for personnel dosimetry. This class is based on the anthropomorphic human phantom and weight factors for effective dose equivalent [14] or effective dose [15].

Computer applications. The 1980s and 1990s were decades of revolution for the computational aspects of radiation shield design and analysis. The advent of inexpensive personal computers with rapidly increasing speeds and memory freed the shielding analyst from dependence on a few supercomputers at national laboratories. Many shielding codes that could previously run only on large mainframe computers were reworked to run on small personal computers, thereby, allowing any shielding analysts to perform detailed calculations that only a privileged few were able to do previously.

At the same time, many improvements were made to the transport codes and their algorithms. MCNP has gone through a series of improvements adding new capabilities and improvements, such as new variance reduction methods, tallies, and physics models. It has also spun off a second version MCNPX with a capability of treating 34 types of particles with energies up to 150 MeV. Also in these decades many other Monte Carlo transport codes were developed by researchers in many nations. Each version has unique features and capabilities. General-purpose discrete-ordinates codes were also extensively improved with many novel acceleration schemes introduced to improve their speeds. An excellent review of many such improvements is given by Adams and Larsen [16].

Practice of Radiation Shielding

Shielding design and shielding analysis are complementary activities. In design, the source and maximum target dose are specified, and the task is to determine the type and amount of the shielding required to reduce the target dose to that specified. In analysis, the source and shielding are identified and the task is to determine the dose at some point(s) of interest. Whether one is engaged in a hand calculation or in a most elaborate Monte Carlo simulation, one is faced with the tasks of (1) characterizing the source, (2) characterizing the nature and attenuating properties of the shielding materials, (3) evaluating at a target location the radiation intensity and perhaps its angular and energy distributions, and (4) converting the intensity to a dose or response meaningful in terms of radiation effects.

Source Characterization

Source geometry, energy, and angular distribution are required characteristics. Radionuclide sources, with isotropic emission and unique energies of gamma and x-rays are relatively easy to characterize. Activity and source strength must be carefully distinguished, as not every decay results in emission of a particular gamma or x-ray. Careful consideration must be given to a low-energy limit below which source particles may be ignored, else computation resources may be wasted. Similarly, when photons of many energies are emitted, as in the case of fission-product sources, one is compelled to use a group structure in source characterization, and much care is needed in establishing efficient and appropriate group energy limits and group average energies. When the source energies are continuously distributed, as is the case with fission neutrons and gamma rays, one option is to use a multigroup approach, as might be used in point-kernel calculations. Another option, useful in Monte Carlo simulations, is to sample source energies from a mathematical representation of the energy spectrum.

A point source is very often an appropriate approximation of a physical source of small size. It is also appropriate to represent a line, plane, or volume source as a collection of point sources, as is done in the point-kernel method of shielding analysis. Radionuclide and fission sources are isotropic in angular distribution; however, there are cases for which it is efficient to identify a surface and to characterize the surface as a secondary source surface. Such surface sources are very often non-isotropic in angular distribution. For example, consider the radiation emitted into the atmosphere from a large body of water containing a distributed radiation source. The interface may be treated approximately, but very effectively, as a plane source emitting radiation not isotropically, but with an intensity varying with the angle of emission from the surface.

Attenuating Properties

The total microscopic cross section for an element or nuclide, $\sigma(E)$, multiplied by the atomic density, is the linear interaction coefficient $\mu(E)$, also called the macroscopic cross section, the probability per unit (differential) path length that a particle of energy E interacts with the medium in some way. Its reciprocal, called the *mean free path*, is the average distance traveled before interaction. Usually, the ratio μ/ρ, called the mass interaction coefficient, is tabulated because it is independent of density. Various subscripts may be used to designate particular types of interactions, for example, $\sigma_{\mathrm{a}}(E)$ for absorption or $\sigma_{\mathrm{f}}(E)$ for fission. Likewise, additional independent variables may be introduced, with, for example, $\sigma_{\mathrm{s}}(E, E')\mathrm{d}E'$ representing the cross section for scattering from energy E to an energy between E' and $E' + \mathrm{d}E'$. Information resources for attenuating properties are described in this entry's historical review, as are resources for radionuclide decay data.

Intensity Characterization

The intensity of a neutron or photon field is usually described in terms of radiation crossing the surface of a small spherical volume V. The *fluence* Φ is defined, in the limit $V \rightarrow 0$, as the expected or average sum of the path lengths in V traveled by entering particles divided by the volume V. Equivalently, Φ is, again in the limit $V \rightarrow 0$, the expected number of particles crossing the surface of V divided by the cross-sectional area of the volume. The time derivative of the fluence is the fluence rate or *flux density* $\dot{\Phi}$. Note that the fluence, though having units of reciprocal area, has no reference area or orientation. Note too that the fluence and flux are point functions. The fluence, a function of position, may also be a distribution function for particle energies and directions. For example, $\Phi(r, E, \Omega)\,\mathrm{d}E\mathrm{d}\Omega$ is the fluence at $\mathbf{r}$ of particles with energies in $\mathrm{d}E$ about E and with directions in solid angle $\mathrm{d}\Omega$ about the direction $\boldsymbol{\Omega}$. When a particular surface, with outward normal $\mathbf{n}$, is used as a reference, it is useful to define radiation intensity in terms of the flow $J_n(\mathbf{r}, E, \boldsymbol{\Omega})\,\mathrm{d}E\mathrm{d}\Omega \equiv \mathbf{n} \cdot \boldsymbol{\Omega}\Phi(\mathbf{r}, E, \boldsymbol{\Omega})\,\mathrm{d}E\mathrm{d}\Omega$ across the reference surface.

Fluence-to-Dose Conversion Factors

Whether the shield designer uses the simplest of the point-kernel methods or the most comprehensive of the Monte Carlo or discrete-ordinates methods, fluence-to-dose conversion factors generally have to be used. The radiation attenuation calculation deals with the particle fluence, the direct measure of radiation intensity. To convert that intensity into a measure of radiation damage or heating of a material, to a field measurement such as *exposure*, or to a measure of health risk, conversion factors must be applied.

The shielding analysis ordinarily yields the energy spectrum $\Phi(\mathbf{r}, E)$ of the photon or neutron fluence at a point identified by the vector $\mathbf{r}$. Use of a Monte Carlo code normally yields the energy spectrum as a function of energy, whence the dose or, more generally, response $R(\mathbf{r})$ is given by the convolution of the fluence with the fluence-to-dose factor, here called the response function $\mathcal{R}(E)$, so that

$$R(\mathbf{r}) = \int_E \mathrm{d}E\mathcal{R}(\mathbf{r}, E)\Phi(\mathbf{r}, E). \tag{1}$$

Point-kernel, or other energy-multigroup methods yield the energy spectrum at discrete energies, or in energy groups, and the dose convolution is a summation rather than an integration.

While the fluence is most always computed as a point function of position, the response of interest may be a dose at a point (called a *local* dose) or it may be a much more complicated function such as the average radiation dose in a physical volume such as an anthropomorphic phantom. Local and phantom-related doses are briefly discussed later.

Suppose the local dose of interest is the kerma, defined as the expected sum of the initial kinetic energies of all charged particles produced by the radiation field in a mass m, in the limit as $m \to 0$. Then the response function is given by

$$\mathcal{R}_{\mathcal{K}}(E) = \kappa \sum_i \frac{N_i}{\rho} \sum_j \sigma_{ji}(E)\varepsilon_{ji}(E). \quad (2)$$

in which ρ is the mass density, N_i is the atoms of species i per unit volume (proportional to ρ), $\sigma_{ji}(E)$ is the cross section for the jth interaction with species i, and $\varepsilon_{ji}(E)$ is the average energy transferred to secondary charged particles in the jth interaction with species i. A units conversion factor κ is needed to convert from, say, units of MeV cm^2/g to units of rad cm^2 or Gy cm^2. For neutrons, a quality factor multiplier $Q(E)$ is needed to convert to units of *dose equivalent* (rem or Sv). For photons, Eq. 2 reduces to

$$\mathcal{R}_{\mathcal{K}}(E)\ (\text{Gy cm}^2) = 1.602 \times 10^{-10} E[\mu_{\text{tr}}(E)/\rho], \quad (3)$$

where E is in MeV and μ_{tr}/ρ is the mass energy transfer coefficient in units of cm^2/g for the material to which energy is transferred.

More related to radiation damage is the local *absorbed dose*, defined as the expected energy imparted, through ionization, excitation, chemical changes, and heat, to a mass m, in the limit as $m \to 0$. Under conditions of *charged-particle equilibrium*, the neutron or gamma-ray kerma equals the absorbed dose, less the energy radiated away as bremsstrahlung. Such equilibrium is approached in a region of homogeneity in composition and uniformity in neutron or photon intensity. Then the absorbed dose is given by Eq. 3 with μ_{tr} replaced by the energy absorption coefficient μ_{en} to account for any bremsstrahlung losses.

The second type of response function or dose is that related to the local dose within a simple geometric phantom or some sort of average dose within an anthropomorphic phantom. The phantom dose, in fact, is a point function and serves as a standardized reference dose for instrument calibration and radiation protection purposes. Even though the radiation fluence, itself a point function, may have strong spatial and angular variation as well as energy variation, it is still possible to associate with the radiation fluence a phantom-related dose. The procedure is as follows. The fluence is treated, for example, as a very broad parallel beam of the same intensity as the actual radiation field, incident in some fixed way on the phantom. This is the so-called expanded and aligned field. For a geometric phantom, the dose is computed at a fixed depth. For an anthropomorphic phantom, the dose is computed as an average of doses to particular tissues and organs, weighted by the susceptibility of the tissues and organs to radiation carcinogenesis or hereditary illness. Many phantoms have been used with various directions of incident radiation. The calculated response functions are then tabulated as a function of the radiation energy. Additional details of phantom doses and their tabulations are given by Shultis and Faw [17].

Basic Analysis Methods

To say modern shielding practice has been reduced to running large "black-box" codes is very misleading. Randomly varying model parameters, such as shield dimensions, placement, and material, is a very inefficient way to optimize shielding for a given situation. Using the concepts and ideas behind the earlier simplified methods often allows a shield analyst to select materials and geometry for a preliminary design before using large transport codes to refine the design. In this section, fundamental methods for estimating neutron or photon doses are reviewed. Such indirectly ionizing radiation is characterized by straight-line trajectories punctuated by "point" interactions. The basic concepts presented here apply equally to all particles of such radiation.

It should be noted that throughout this entry, calculated doses are the *expected* or *average* value of the stochastic measured doses, that is, the mechanistically calculated dose represents the statistical average of a large number of dose measurements which exhibit random fluctuations as a consequence of the stochastic nature of the source emission and interactions in the detector and surrounding material.

Uncollided Radiation Doses

In many situations, the dose at some point of interest is dominated by particles streaming directly from the source without interacting in the surrounding medium. For example, if only air separates a gamma-ray or neutron source from a detector, interactions in the intervening air or in nearby solid objects, such as the ground or building walls, are often negligible, and the radiation field at the detector is due almost entirely to uncollided radiation coming directly from the source.

In an attenuating medium, the uncollided dose at a distance r from a point isotropic source emitting S_p particles of energy E is

$$D^\text{o}(r) = \frac{S_\text{p}\mathcal{R}}{4\pi r^2}\,\text{e}^{-l}, \tag{4}$$

where l is the total number of mean-free-path lengths of material a particle must traverse before reaching the detector, namely, $\int_0^r \text{d}s\,\mu(s)$. Here $\mathcal{R}$ is the appropriate response function. The $1/(4\pi r^2)$ term in Eq. 4 is often referred to as the *geometric* attenuation and the e^{-l} term the *material* attenuation. Equation 4 can be extended easily to a source emitting particles with different discrete energies or a continuous spectrum of energies.

Point Kernel for Uncollided Dose

Consider an isotropic point source placed at $\mathbf{r}_\text{s}$ and an isotropic point detector (or *target*) placed at $\mathbf{r}_\text{t}$ in a homogeneous medium. The detector response depends not on $\mathbf{r}_\text{s}$ and $\mathbf{r}_\text{t}$ separately, but only on the distance $|\mathbf{r}_\text{s} - \mathbf{r}_\text{t}|$ between the source and detector. For a unit strength source, the detector response is (cf. Eq. 4)

$$\mathcal{G}^\text{o}(\mathbf{r}_\text{s}, \mathbf{r}_\text{t}, E) = \frac{\mathcal{R}(E)}{4\pi|\mathbf{r}_\text{s} - \mathbf{r}_\text{t}|^2}\,\text{e}^{-\mu(E)|\mathbf{r}_\text{s}-\mathbf{r}_\text{t}|}. \tag{5}$$

Here $\mathcal{G}^\text{o}(\mathbf{r}_\text{s}, \mathbf{r}_\text{t}, E)$ is the *uncollided dose point kernel* and equals the dose at $\mathbf{r}_\text{t}$ per particle of energy E emitted isotropically at $\mathbf{r}_\text{s}$. This result holds for any geometry or medium provided that the material through which a ray from $\mathbf{r}_\text{s}$ to $\mathbf{r}_\text{t}$ passes has a constant interaction coefficient μ.

With this point kernel, the uncollided dose due to an arbitrarily distributed source can be found by first decomposing (conceptually) the source into a set of contiguous effective point sources and then summing (integrating) the dose produced by each point source.

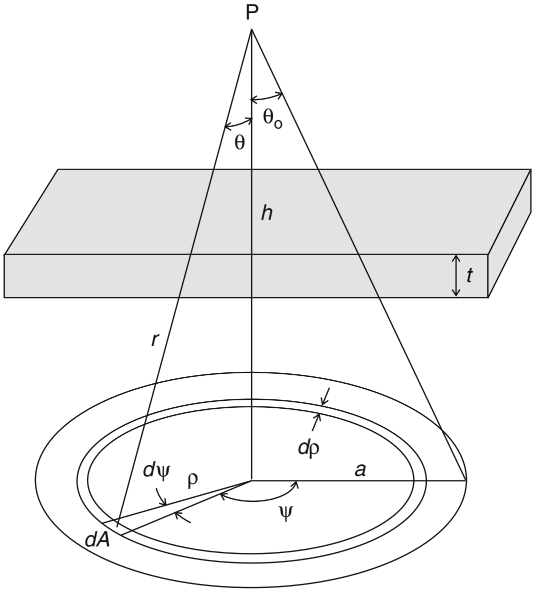
Applications to Selected Geometries

The results for the uncollided dose from a point source can be used to derive expressions for the uncollided dose arising from a wide variety of distributed sources such as line sources, area sources, and volumetric sources [4, 5, 18, 19]. An example to illustrate the method is as follows:

An isotropic disk source of radius a emitting isotropically S_a particles per unit area at energy E is depicted in Fig. 1. A detector is positioned at point P a distance h above the center of the disk. Suppose the only material separating the disk source and the receptor at P is a slab of thickness t with a total attenuation coefficient μ.

Consider a differential area dA between distance ρ and $\rho + \text{d}\rho$ from the disk center and between ψ and $\psi + \text{d}\psi$. The source within dA may be treated as an effective point isotropic source emitting $S_a\text{d}A = S_a\rho\text{d}\rho\text{d}\psi$ particles which produces an uncollided dose at P of dD^o. The ray from the source in dA must pass through a slant distance of the shield $t\sec\theta$ so that the dose at P from particles emitted in dρ about ρ is

$$\text{d}D^\text{o}(P) = \frac{\mathcal{R}S_a\rho\text{d}\rho\text{d}\psi}{4\pi r^2}\exp[-\mu t\sec\theta], \tag{6}$$



Radiation Shielding. Figure 1
An isotropic disk source is shielded by a parallel slab shield of thickness *t*

where $\mathcal{R}$ and μ generally depend on the particle energy E. To obtain the total dose at P from all differential areas of the disk source, one then must sum, or rather integrate, $\mathrm{d}D^{\mathrm{o}}$ over all differential areas. Thus, the total uncollided dose at P is

$$D^{\mathrm{o}}(P) = \frac{S_a\mathcal{R}}{4\pi}\int_0^{2\pi}\mathrm{d}\psi\int_0^a \mathrm{d}\rho\,\frac{\rho exp[-\mu t\sec\theta]}{r^2}. \tag{7}$$

Because h is fixed, $\rho\mathrm{d}\rho = r\mathrm{d}r$, and from Fig. 1 it is seen that $r = h\sec\theta$. Integration over ψ and changing variables yields

$$D^{\mathrm{o}}(P) = \frac{S_a\mathcal{R}}{2}\int_h^{h\sec\theta_{\mathrm{o}}}\mathrm{d}r\,r^{-1}\mathrm{e}^{-\mu rt/h} \tag{8}$$

$$= \frac{S_a\mathcal{R}}{2}\int_{\mu t}^{\mu t\sec\theta_o}\mathrm{d}x\,x^{-1}\mathrm{e}^{-x} \tag{9}$$

$$= \frac{S_a\mathcal{R}}{2}[E_1(\mu t) - E_1(\mu t\sec\theta_{\mathrm{o}})], \tag{10}$$

where the *exponential integral function* E_n is defined as $E_n(x) \equiv x^{n-1}\int_x^\infty \mathrm{d}u\,u^{-n}\mathrm{e}^{-u}$ and is tabulated in many compilations [4, 5, 17].

Intermediate Methods for Photon Shielding

In this section, several special techniques are summarized for the design and analysis of shielding for gamma and x-rays with energies from about 1 keV to about 20 MeV. These techniques are founded on very precise radiation transport calculations for a wide range of carefully prescribed situations. These techniques, which rely on buildup factors, attenuation factors, albedos or reflection factors, and line-beam response functions, then allow estimation of photon doses for many frequently encountered shielding situations without the need of transport calculations.

Buildup-Factor Concept

The total photon fluence $\Phi(\mathbf{r},E)$ at some point of interest $\mathbf{r}$ is the sum of two components: the *uncollided* fluence $\Phi^{\mathrm{o}}(\mathbf{r},E)$ of photons that have streamed to $\mathbf{r}$ directly from the source without interaction, and the fluence of *scattered* and *secondary* photons $\Phi^{\mathrm{s}}(\mathbf{r},E)$ consisting of source photons scattered one or more times, as well as secondary photons such as x-rays and annihilation gamma rays.

The buildup factor $B(\mathbf{r})$ is defined as

$$B(\mathbf{r}) \equiv \frac{D(\mathbf{r})}{D^{\mathrm{o}}(\mathbf{r})} = 1 + \frac{D^{\mathrm{s}}(\mathbf{r})}{D^{\mathrm{o}}(\mathbf{r})}, \tag{11}$$

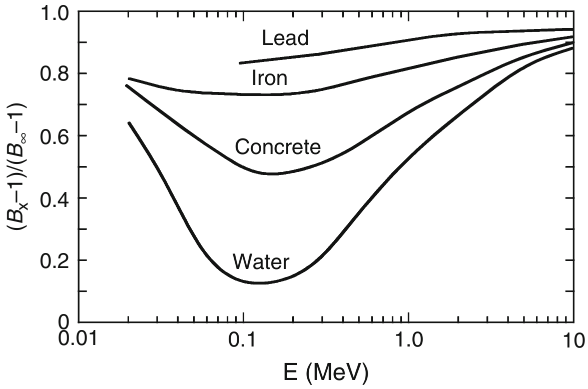
where $D(\mathbf{r})$ is the total dose equal to the sum of the uncollided dose $D^{\mathrm{o}}(\mathbf{r})$ and the scattered or secondary photon dose $D^{\mathrm{s}}(\mathbf{r})$. For a monoenergetic source this reduces to

$$B(E_{\mathrm{o}},\mathbf{r}) = 1 + \frac{1}{\Phi^{\mathrm{o}}(\mathbf{r})}\int_0^{E_{\mathrm{o}}}\mathrm{d}E\,\frac{\mathcal{R}(E)}{\mathcal{R}(E_{\mathrm{o}})}\Phi^{\mathrm{s}}(\mathbf{r},E). \tag{12}$$

In this case, the nature of the dose or response is fully accounted for in the ratio $\mathcal{R}(E)/\mathcal{R}(E_{\mathrm{o}})$. By far the largest body of buildup-factor data is for point, isotropic, and monoenergetic sources of photons in infinite homogeneous media. Calculation of buildup factors for high-energy photons requires consideration of the paths traveled by positrons from their creation until their annihilation. Such calculations have been performed by Hirayama [20] and by Faw and Shultis [21] for photon energies as great as 100 MeV. Because incoherent scattering was neglected in many buildup-factor calculations, coherent scattering should also be neglected in calculating the uncollided dose, a significant consideration only for low-energy photons at deep penetration.

Buildup-Factor Geometry Generally, buildup factors depend on the source and shield geometries. For a given material thickness between source and detector, buildup factors are slightly different for point isotropic sources in (a) an infinite medium, (b) at the surface of a bare sphere, and for a slab shield between source and detector. However, the use of buildup factors for a point isotropic source is almost always conservative, that is, the estimated dose is greater than that for a finite shield [17]. Adjustment factors for buildup factors at the surface of a finite medium in terms of the infinite-medium buildup factors is illustrated in Fig. 2.

Buildup factors are also available for plane isotopic (PLI) and plane monodirectional (PLM) gamma-ray sources in infinite media. Indeed, Fano et al. [22], Goldstein [23], and Spencer [24], in their moments-method calculations, obtained buildup factors for plane sources first and, from these, buildup factors for point sources



Radiation Shielding. Figure 2
Adjustment factor for the buildup factor B_x at the boundary of a finite medium in terms of the infinite-medium buildup factor B_∞ for the same depth of penetration (EGS4 calculations courtesy of Sherrill Shue, Nuclear Engineering Department, Kansas State University)

were derived. Buildup factors at depth in a half-space shield are also available for the PLM source, that is, normally incident photons [20, 25, 26]. The use of buildup factors for a point isotropic source in an infinite medium is conservative, that is, overpredictive, for the PLI and PLM geometries.

Buildup Factors for Stratified Shields Sometimes shields are stratified, that is, composed of layers of different materials. The use of the buildup-factor concept for such heterogeneous shields is, for the most part, of dubious merit. Nevertheless, implementation of point-kernel codes for shielding design and analysis demands some way of treating buildup when the ray from source point to dose point is through more than one shielding material. However, certain regularities do exist, which permit approximate use of homogeneous-medium buildup factors for stratified shields. Many approximate buildup methods have been suggested, as described by Shultis and Faw [17]; however, they are of little use in most point-kernel codes and are not needed at all for shielding analysis based on transport methods.

Point-Kernel Computer Codes

There are many codes in wide use that are based on the point-kernel technique. In these codes, a distributed source is decomposed into small but finite elements and the dose at some receptor point from each element is computed using the uncollided dose kernel and a buildup factor based on the optical thickness of material between the source element and the receptor. The results for all the source elements are then added together to obtain the total dose. Some that have been widely used are MicroShield [27], the QAD series [28], QADMOD-GP [29], QAD-CGGP [30], and G^3 [31].

Broad-Beam Attenuation

Often a point radionuclide or x-ray source in air is located sufficiently far from a wall or shielding slab that the radiation reaches the wall in nearly parallel rays. Further, the attenuation in the air is usually quite negligible in comparison to that provided by the shielding wall. Shielding design and analysis for such broad-beam illumination of a slab shield are addressed by NCRP Report 49 [9], Archer [32], and Simpkin [33]. The dose at the surface of the cold side of the wall can be computed as

$$D = D^{\mathrm{o}} A_{\mathrm{f}}. \tag{13}$$

For a radionuclide source of activity $\mathcal{A}$, the dose D^{o} without the wall can be expressed in terms of the source energy spectrum, response functions, and distance r from the source to the cold side of the wall. Then,

$$D = D^{\mathrm{o}} A_{\mathrm{f}} = \frac{\mathcal{A}}{r^2} \Gamma A_{\mathrm{f}}, \tag{14}$$

where Γ, called the *specific gamma-ray constant*, is the dose rate in vacuum at a unit distance from a source with unit activity, and A_{f} is an *attenuation factor* which depends on the nature and thickness of the shielding material, the source energy characteristics, and the angle of incidence θ (with respect to the wall normal). Values for Γ and A_{f} are provided by NCRP [9].

Oblique Incidence Attenuation factors for obliquely incident beams are presented in NCRP Report 49 [9]. For such cases, special three-argument slant-incidence buildup factors should be used [17]. For a shield wall of thickness t mean free paths, slant incidence at angle θ with respect to the normal to the wall, and source energy E_{o}, the attenuation factor is in function form $A_{\mathrm{f}}(E_{\mathrm{o}}, t, \theta)$. However, a common, but erroneous, practice has been to use a two-argument attenuation factor based on an infinite-medium buildup factor for slant penetration distance $t\sec\theta$, in the form $A_f(E_o, t\sec\theta)$. This practice can lead to severe underprediction of transmitted radiation doses.

X-ray Beam Attenuation For x-ray sources, the appropriate measure of source strength is the electron-beam current i, and the appropriate characterization of photon energies, in principle, involves the peak accelerating voltage (kVp), the wave form, and the degree of filtration (e.g., beam half-value thickness). If i is the beam current (mA) and r is the source-detector distance (m), the dose behind a broadly illuminated shield wall is

$$D(P) = \frac{i}{r^2} K_o A_f, \tag{15}$$

in which K_o, called the *radiation output (factor)*, is the dose rate in vacuum (or air) per unit beam current at unit distance from the source in the absence of the shield. Empirical formulas for computing A_f are available for shield design [34, 35].

Intermediate Methods for Neutron Shielding

Shielding design for fast neutrons is generally far more complex than shielding design for photons. Not only does one have to protect against the neutrons emitted by some source, one also needs to protect against primary gamma rays emitted by most neutron sources as well as secondary photons produced by inelastic neutron scattering and from radiative capture. There may also be secondary neutrons produced from $(n, 2n)$ and fission reactions. In many instances, secondary photons produce greater radiological risks than do the primary neutrons. Fast-neutron sources include spontaneous and induced fission, fusion, (α, n) reactions, (γ, n) reactions, and spallation reactions in accelerators, each producing neutrons with a different distribution of energies.

Unlike photon cross sections, neutron cross sections usually vary greatly with neutron energy and among the different isotopes of the same element. Comprehensive cross-section databases are needed. Also, because of the erratic variation of the cross sections with energy, it is difficult to calculate uncollided doses needed in order to use the buildup-factor approach. Moreover, buildup factors are very geometry dependent and sensitive to the energy spectrum of the neutron fluence and, consequently, point-kernel methods can be applied to neutron shielding only in very limited circumstances.

Early work led to kernels for fission sources in aqueous systems and the use of removal cross sections to account for shielding barriers. Over the years, the methodology was stretched to apply to nonaqueous hydrogenous media, then to non-hydrogenous media, then to fast-neutron sources other than fission. Elements of diffusion and age theory were melded with the point kernels. Today, with the availability of massive computer resources, neutron shielding design and analysis is largely done using transport methods. Nevertheless, the earlier methodologies offer insight and allow more critical interpretation of transport calculations.

Also, unlike ratios of different photon response functions, those for neutrons vary, often strongly, with neutron energy. Hence, neutrons doses cannot be converted to different dose units by simply multiplying by an appropriate constant. The energy spectrum of the neutron fluence is needed to obtain doses in different units. Consequently, many old measurements or calculations of point kernels, albedo functions, transmission factors, etc., made with obsolete dose units cannot be converted to modern units because the energy spectrum is unknown. In this case, there is no recourse but to repeat the measurements or calculations.

Capture Gamma Photons

A significant, often dominant, component of the total dose at the surface of a shield accrues from capture gamma photons produced deep within the shield and arising from neutron absorption. Of lesser significance are secondary photons produced in the inelastic scattering of fast neutrons. Secondary neutrons are also produced as a result of (γ, n) reactions. Thus, in transport methods, gamma-ray and neutron transport are almost always coupled.

Historically, capture gamma-ray analysis was appended to neutron removal calculations. Most neutrons are absorbed only when they reach thermal energies, and, consequently, only the absorption of thermal neutrons was considered. (Exceptional cases include the strong absorption of epithermal neutrons in fast reactor cores or in thick slabs of low-moderating, high-absorbing material.) For this reason, it is important to calculate accurately the thermal neutron fluence $\Phi_{\mathrm{th}}(\mathbf{r})$ in the shield. The volumetric source strength of capture photons per unit energy about E is then given by

$$S_\gamma(\mathbf{r}, E) = \Phi_{\mathrm{th}}(\mathbf{r})\mu_\gamma(\mathbf{r})f(\mathbf{r}, E), \tag{16}$$

where $\mu_\gamma(\mathbf{r})$ is the absorption coefficient at $\mathbf{r}$ for thermal neutrons and $f(\mathbf{r}, E)$ is the number of photons produced in unit energy about E per thermal neutron absorption at $\mathbf{r}$.

Once the capture gamma-ray source term $S_\gamma(\mathbf{r}, E)$ is known throughout the shield, point-kernel techniques using exponential attenuation and buildup factors can be used to calculate the capture gamma-ray dose at the shield surface.

Neutron Shielding with Concrete

Concrete is probably the most widely used shielding material because of its relatively low cost and the ease with which it can be cast into large and variously shaped shields. However, unlike that for photon attenuation in concrete, the concrete composition, especially the water content, has a strong influence on its neutron attenuation properties. Other important factors that influence the effectiveness of concrete as a neutron shield include type of aggregate, the dose–response function, and the angle of incidence of the neutrons.

Because concrete is so widely used as a shield material, its effectiveness for a monoenergetic, broad, parallel beam of incident neutrons has been extensively studied, both for normal and slant incidence, and many tabulated results for shields of various thickness are available [36–40]. These results, incorporated into design and manufacturing standards (standards are available from professional societies such as the American Nuclear Society and the American Society of Mechanical Engineers) are extremely useful in the preliminary design of concrete shields.

Gamma-Ray and Neutron Reflection

Until now only shielding situations have been considered in which the radiation reaching a target contains an uncollided component. For these situations, point-kernel approximations, in principle, may be used and concepts such as particle buildup may be applied. However, in many problems encountered in shielding design and analysis, only scattered radiation may reach the target. Radiation doses due to reflection from a surface are examples that arise in treatment of streaming of radiation through multilegged ducts and passageways. Treatment of radiation reflection from surfaces of structures is also a necessary adjunct to precise calibration of nuclear instrumentation. Skyshine, that is, reflection in the atmosphere of radiation from fixed sources to distant points is another example of this class of reflected-radiation problems. All such reflection problems are impossible to treat using elementary point-kernel methods and are also very difficult and inefficient to treat using transport-based methods. For reflection from a surface of radiation from a point source to a point receiver, the *albedo function* has come to be very useful in design and analysis. The same can be said for use of the *line-beam response function* in treatment of skyshine. Both are discussed below.

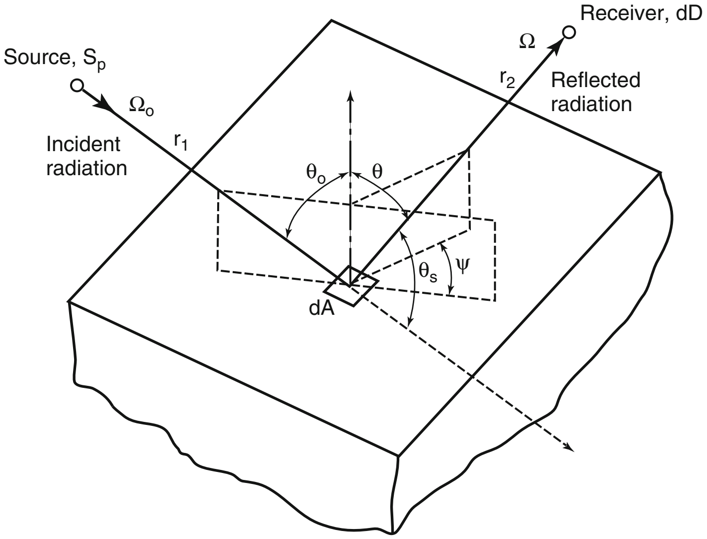
Albedo Methods

There are frequent instances for which the dose at some location from radiation reflected from walls and floors may be comparable to the line-of-sight dose. The term *reflection* in this context does not imply a surface scattering. Rather, gamma rays or neutrons penetrate the surface of a shielding or structural material, scatter within the material, and then emerge from the material with reduced energy and at some location other than the point of entry.

In many such analyses, a simplified method, called the *albedo method*, may be used. The albedo method is based on the following approximations. (1) The displacement between points of entry and emergence may be neglected. (2) The reflecting medium is effectively a half-space, a conservative approximation. (3) Scattering in air between a source and the reflecting surface and between the reflecting surface and the detector may be neglected.

Application of the Albedo Method

Radiation reflection may be described in terms of the geometry shown in Fig. 3. Suppose that a point isotropic and monoenergetic source is located distance r_1



Radiation Shielding. Figure 3
Angular and energy relationships in the albedo formulation

from area dA along incident direction $\mathbf{\Omega}_o$ and that a dose point is located distance r_2 from area dA along emergent direction $\mathbf{\Omega}$. Suppose also the source has an angular distribution such that $S(\theta_o)$ is the source intensity per steradian evaluated at the direction from the source to the reflecting area dA. Then the dose dD_r at the detector from particles reflected from dA can be shown to be [17]

$$\mathrm{d}D_r = D_o \alpha_D(E_o, \theta_o; \theta, \psi) \frac{\mathrm{d}A \cos\theta_o}{r_2^2}, \tag{17}$$

in which D_o is the dose at dA due to incident particles. Here $\alpha_D(E_o, \theta_o; \theta, \psi)$ is the *dose albedo*. Determination of the total reflected dose D_r requires integration over the area of the reflecting surface. In doing so one must be aware that, as the location on the surface changes, all the variables $\theta_o, \theta, \psi, r_1$, and r_2 change as well. Also, it is necessary to know $\alpha_D(E_o, \theta_o; \theta, \psi)$ or, more usefully, to have some analytical approximation for the dose albedo so that numerical integration over all the surface area can be performed efficiently.

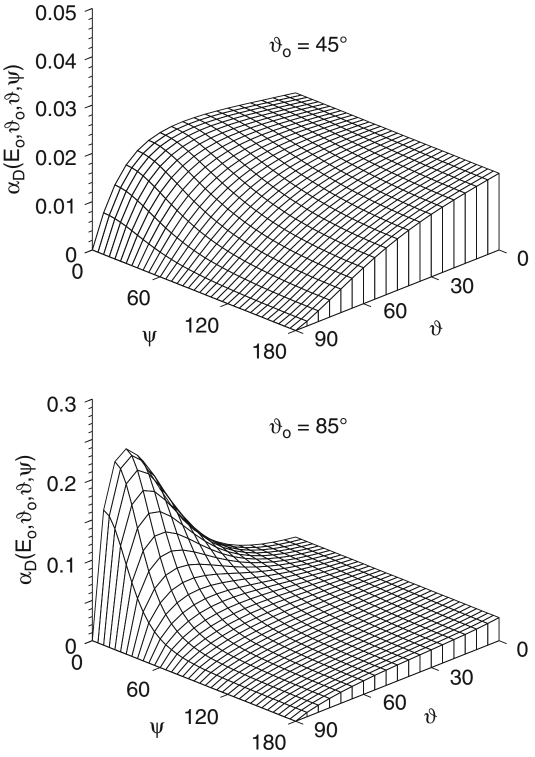
Gamma-Ray Dose Albedo Approximations

A two-parameter approximation for the photon dose albedo was first devised by Chilton and Huddleston [41] and later extended by Chilton et al. [42]. Chilton [43] later proposed a more accurate seven-parameter albedo formula for concrete. Brockhoff [44] published seven-parameter fit data for albedos from water, concrete, iron, and lead. Two examples of this dose albedo approximation are shown in Fig. 4.

Neutron Dose Albedo Approximations

The dose albedo concept is very useful for streaming problems that involve "reflection" of neutrons or photons from some material interface. However, unlike photon albedos, the neutron albedos are seldom tabulated or approximated for monoenergetic incident neutrons because of the rapid variation with energy of neutron cross sections. Rather, albedos for neutrons with a specific range of energies (energy group) are usually considered, thereby, averaging over all the cross-section resonances in the group. Also unlike photon albedos, neutron albedos involve reflected dose from both neutrons and secondary capture gamma rays.

There are many studies of the neutron albedos in the literature. Selph [45] published a detailed review. Extensive compilations of neutron albedo data are available, for example, SAIL [46] and BREESE-II [47]. Of more utility are analytic approximations for the albedo based on measured or calculated albedos. Neutrons albedos are often divided into three types: (1) fast-neutron albedos ($E \geq 0.2$ MeV), (2) intermediate-energy albedos, and (3) thermal-neutron albedos. Selph [45] reviews early



Radiation Shielding. Figure 4
Ambient-dose-equivalent albedos for reflection of 1.25-MeV photons from concrete, computed using the seven-term Chilton–Huddleston approximation

approximations for neutron albedos, among which is a 24-parameter approximation developed by Maerker and Muckenthaler [48]. Newly computed and more accurate fast-neutron albedos, based on different 24-parameter approximations, have been computed by Brockhoff [44] for several shielding materials.

For neutrons with energy less than about 100 keV, the various dose equivalent response functions are very insensitive to neutron energy. Consequently, the dose albedo α_D is very closely approximated by the number albedo α_N. Thus, for reflected dose calculations involving intermediate or thermal neutrons, the number albedo is almost always used. Coleman et al. [49] calculated neutron albedos for intermediate-energy neutrons (200 keV to 0.5 eV) incident monodirectionally on reinforced concrete slabs and developed a nine-parameter formula for the albedo.

Thermal neutrons entering a shield undergo isotropic scattering that, on the average, does not change their energies. For one-speed particles incident in an azimuthally symmetric fashion on a half-space of material that isotropically scatters particles, Chandrasekhar [50] derived an exact expression for the differential albedo. A purely empirical and particularly simple formula, based on Monte Carlo data for thermal neutrons, has been proposed by Wells [51] for ordinary concrete, namely,

$$\alpha_N(\theta_o;\theta,\psi) = 0.21\cos\theta(\cos\theta_o)^{-1/3}. \qquad (18)$$

Radiation Streaming Through Ducts

Except in the simplest cases, the analysis of radiation streaming requires advanced computational procedures. However, even within the framework of Monte Carlo transport calculations, albedo methods are commonly used, and special data sets have been developed for such use [46, 47, 52].

Elementary methods for gamma-ray streaming are limited to straight cylindrical ducts, with incident radiation symmetric about the duct axis and uniform over the duct entrance. Transmitted radiation generally may be subdivided into three components: line-of-sight, lip-penetrated, and wall scattered. The first two may be treated using point-kernel methodology. The last requires use of albedo methods to account for scattering over the entire surface area of the duct walls. Selph [45] reviews the methodology of duct transmission calculations and LeDoux and Chilton [53] devised a method of treating two-legged rectangular ducts, important in analysis of structure shielding.

Neutron streaming through gaps and ducts in a shield is much more serious for neutrons than for gamma photons. Neutron albedos, especially for thermal neutrons, are generally much higher than those for photons, and multiple scattering within the duct is very important. Placing bends in a duct, which is very effective for reducing gamma-ray penetration, is far less effective for neutrons. Fast neutrons entering a duct in a concrete shield become thermalized and thereafter are capable of scattering many times, allowing the neutrons to stream through the duct, even those with several bends. Also, unlike gamma-ray streaming, the duct need not be a void (or gas filled) but can be any part of a heterogeneous shield that is "transparent" to neutrons. For example, the steel walls of a water pipe embedded in a concrete shield (such as the cooling pipes that penetrate the biological shield of a nuclear reactor) act as an annular duct for fast neutrons.

There is much literature on experimental and calculational studies of gamma-ray and neutron streaming through ducts. In many of these studies, empirical formulas, obtained by fits to the data, have been proposed. These formulas are often useful for estimating duct-transmitted doses under similar circumstances. As a starting point for finding such information, the interested reader is referred to Rockwell [18], Selph [45], and NCRP [54].

Gamma-Ray and Neutron Skyshine

For many intense localized sources of radiation, the shielding against radiation that is directed skyward is usually far less than that for the radiation emitted laterally. However, the radiation emitted vertically into the air undergoes scattering interactions and some radiation is reflected back to the ground, often at distances far from the original source. This atmospherically reflected radiation, referred to as *skyshine*, is of concern both to workers at a facility and to the general population outside the facility site.

As alternatives to rigorous transport-theory treatment of the skyshine problem several approximate procedures have been developed for both gamma-photon and neutron skyshine sources [54]. This section summarizes one approximate method, which has been found useful for bare or shielded skyshine sources. The *integral line-beam skyshine method*, is based on the availability of a *line-beam response function* $\mathcal{R}(E, \phi, x)$, which gives the dose (air kerma or ambient dose) at a distance x from a point source emitting a photon or neutron of energy E at an angle ϕ from the source-to-detector axis into an infinite air medium. The air–ground interface is neglected in this method. This response function can be fit over a large range of x to the following three-parameter empirical formula, for a fixed value of E and ϕ [55]:

$$\mathcal{R}(E, \phi, x) = \kappa(\rho/\rho_o)^2 E[x(\rho/\rho_o)]^b \exp[a - cx(\rho/\rho_o)], \tag{19}$$

in which ρ is the air density in the same units as the reference density $\rho_o = 0.0012\ \mathrm{g/cm^3}$. The constant κ depends on the choice of units.

The parameters a, b, and c in Eq. 19 depend on the photon or neutron energy and the source emission angle. These parameters have been estimated and tabulated, for fixed values of E and ϕ, by fitting Eq. 19 to values of the line-beam response function, at different x distances, usually obtained by Monte Carlo calculations. Gamma-ray response functions have been published by Lampley [56] and Brockhoff et al. [57]. Neutron and secondary gamma-ray response functions have been published by Lampley [56] and Gui et al. [58]. These data and their method of application are presented by Shultis and Faw [17].

To obtain the skyshine dose $D(d)$ at a distance d from a bare collimated source, the line-beam response function, weighted by the energy and angular distribution of the source, is integrated over all source energies and emission directions. Thus, if the collimated source emits $S(E, \boldsymbol{\Omega})$ photons, the skyshine dose is

$$D(d) = \int_0^\infty \mathrm{d}E \int_{\Omega_s} \mathrm{d}\Omega\, S(E, \boldsymbol{\Omega})\mathcal{R}(E, \phi, d), \tag{20}$$

where the angular integration is over all emission directions $\boldsymbol{\Omega}_s$ allowed by the source collimation. Here, ϕ is a function of the emission direction $\boldsymbol{\Omega}$. To obtain this result, it has been assumed that the presence of an air–ground interface can be neglected by replacing the ground by an infinite air medium. The effect of the ground interface on the skyshine radiation, except at positions very near a broadly collimated source, has been found to be very small.

The presence of a shield over a skyshine source, for example, a building roof, causes some of the source particles penetrating the shield to be degraded in energy and angularly redirected before being transported through the atmosphere. The effect of an overhead shield on the skyshine dose far from the source can be accurately treated by a two-step hybrid method [59,60]. First a transport calculation is performed to determine the energy and angular distribution of the radiation penetrating the shield, and then, with this distribution as an effective point, bare, skyshine source, the integral line-beam method is used to evaluate the skyshine dose.

The integral line-beam method for gamma-ray and neutron skyshine calculations has been applied to a variety of source configurations and found to give generally excellent agreement with benchmark calculations and experimental results [59]. It has been used as the basis of the microcomputer code *MicroSkyshine* [61] for gamma rays. A code package for both neutron and gamma-ray calculations is available from the Radiation Safety Computation Information Center. (Code package CCC-646: SKYSHINE-KSU: Code System to Calculate Neutron and Gamma-Ray Skyshine Doses Using the Integral Line-Beam Method, and data library DLC-188: SKYDATA-KSU: Parameters for Approximate Neutron and Gamma-Ray Skyshine Response Functions and Ground Correction Factors.)

Transport Theory

For difficult shielding problems in which simplified techniques such as point kernels with buildup corrections cannot be used, calculations based on transport theory must often be used. There are two basic approaches for transport calculations: *deterministic* transport calculations in which the linear Boltzmann equation is solved numerically, and *Monte Carlo* calculations in which a simulation is made of how particles migrate stochastically through the problem geometry. Both approaches have their advantages and weaknesses. Because of space limitations, it is not possible to give a detailed review of the vast literature supporting both approaches. What follows is a brief explanation of the basic ideas involved and some general references are supplied.

Deterministic Transport Theory

The neutron or photon flux $\phi(\mathbf{r}, E, \boldsymbol{\Omega})$ for particles with energy E and direction $\boldsymbol{\Omega}$ is rigorously given by the linear Boltzmann equation or, simply, the transport equation

$$\boldsymbol{\Omega} \cdot \nabla\phi(\mathbf{r}, E, \boldsymbol{\Omega}) + \mu(\mathbf{r}, E)\phi(\mathbf{r}, E, \boldsymbol{\Omega}) = S(\mathbf{r}, E, \boldsymbol{\Omega}) + \int_0^\infty \mathrm{d}E' \int_{4\pi} \mathrm{d}\boldsymbol{\Omega}'\, \mu_s(\mathbf{r}, E', \boldsymbol{\Omega}' \rightarrow E, \boldsymbol{\Omega})\phi(\mathbf{r}, E', \boldsymbol{\Omega}'), \tag{21}$$

where S is the volumetric source strength of particles. This equation can be formally integrated to yield the integral form of the transport equation, namely,

$$\phi(\mathbf{r}, E, \boldsymbol{\Omega}) = \phi(\mathbf{r} - R\boldsymbol{\Omega}, E, \boldsymbol{\Omega})f(R) + \int_0^R \mathrm{d}R' q(\mathbf{r} - R'\boldsymbol{\Omega}, \boldsymbol{\Omega})f(R'), \tag{22}$$

where $f(x) \equiv \exp\left[-\int_0^x \mu(\mathbf{r} - R''\boldsymbol{\Omega}, E)\,\mathrm{d}R''\right]$ and q is given by

$$q(\mathbf{r}, E, \boldsymbol{\Omega}) \equiv S(\mathbf{r}, E, \boldsymbol{\Omega}) + \int_0^\infty \mathrm{d}E' \int_{4\pi} \mathrm{d}\boldsymbol{\Omega}' \mu_s(\mathbf{r}, E', \boldsymbol{\Omega}' \rightarrow E, \boldsymbol{\Omega})\phi(\mathbf{r}, E', \boldsymbol{\Omega}'). \tag{23}$$

Unfortunately, neither of these formulations of the transport equation can be solved analytically except for idealistic cases, for example, infinite medium with monoenergetic particles or a purely absorbing medium. Numerical solutions must be used for all practical shielding analyses. Many approximations of the transport equation are used, such as diffusion theory, to allow easier calculations. Also the energy region of interest is usually divided into a few or even hundreds of contiguous energy subintervals and average cross sections are calculated for each group using an assumed energy spectrum of the radiation. In this manner, the transport equation is approximated by a set of coupled equations in which energy is no longer an independent variable. Even with an energy-multigroup approximation, numerical solutions are still computationally formidable.

The most widely used deterministic transport approach is the discrete-ordinates method. In this method, a spatial and directional mesh is created for the problem geometry, and the multigroup form of the transport equation is then integrated over each spatial and directional cell. The solution of the approximating algebraic equations is then accomplished by introducing another approximation that relates the cell-centered flux densities to those on the cell boundaries, and an iterative procedure between the source (scattered particles and true source particles) and flux density calculation is then used to calculate the fluxes at the mesh nodes. For details of this method, the reader is referred to Carlson and Lathrop [62], Duderstadt and Martin [63], and Lewis and Miller [64].

Discrete-ordinates calculations can be computationally expensive because of the usually enormous number of mesh nodes and the fact that the convergence of an iterative solution is often very slow. A subject of great interest in the last 30 years has been the development of numerous methods to accelerate convergence of the iterations. Without convergence acceleration schemes, discrete-ordinate solutions would be computationally impractical for many shielding problems. An excellent description of the various acceleration schemes that have been used is provided by Adams and Larsen [16].

Mature computer codes based on the discrete-ordinates method are widely available to treat one-, two-, and three-dimensional problems in the three basic geometries (rectangular, spherical, and cylindrical) with an arbitrary number of energy groups [65,66].

Although discrete-ordinates methods are widely used by shielding analysts, these methods do have their limitations. Most restrictive is the requirement that the problem geometry must be one of the three basic geometries (rectangular, spherical, or cylindrical) with boundaries and material interfaces placed perpendicular to a coordinate axis. Problems with irregular boundaries and material distributions are difficult to solve accurately with the discrete-ordinates method. Also, in multidimensional geometries, the discrete-ordinates method often produces spurious oscillations in the flux densities (the *ray effect*) as an inherent consequence of the angular discretization. Finally, the discretization of the spatial and angular variables introduces numerical truncation errors, and it is necessary to use sufficiently fine angular and spatial meshes to obtain flux densities that are independent of the mesh size. For multidimensional situations in which the flux density is very anisotropic in direction and in which the medium is many mean-free-path lengths in size, typical of many shielding problems, the computational effort to obtain an accurate discrete-ordinates solution can become very large. However, unlike Monte Carlo calculations, discrete-ordinates methods can treat very-deep-penetration problems, that is, the calculation of fluxes and doses at distances many mean-free-path lengths from a source.

Monte Carlo Transport Theory

In Monte Carlo calculations, particle tracks are generated by simulating the stochastic nature of the particle interactions with the medium. One does not even need to invoke the transport equation; all one needs are complete mathematical expressions of the probability relationships that govern the track length of an individual particle between interaction points, the choice of an interaction type at each such point, the choice of a new energy and a new direction if the interaction is of a scattering type, and the possible production of additional particles. These are all stochastic variables, and in order to make selections of specific values for these variables, one needs a complete understanding of the various processes a particle undergoes in its lifetime from the time it is given birth by the source until it is either absorbed or leaves the system under consideration.

The experience a particle undergoes from the time it leaves its source until it is absorbed or leaves the system is called its *history*. From such histories expected or average values about the radiation field can be estimated. For example, suppose the expected energy $\langle E\rangle$ absorbed in some small volume V in the problem geometry is being sought. There is a probability $f(E)\mathrm{d}E$ that a particle deposits energy in $\mathrm{d}E$ about E. Then the expected energy deposited is simply $\langle E\rangle = \int Ef(E)\mathrm{d}E$. Unfortunately, $f(E)$ is not known *a priori* and must be obtained from a transport calculation. In a Monte Carlo analysis, $f(E)$ is constructed by *scoring* or tallying the energy deposited E_i in V by the ith particle history. Then in the limit of a large number of histories N

$$\langle E\rangle \equiv \int Ef(E)\mathrm{d}E \simeq \overline{E} \equiv \frac{1}{N}\sum_{i=1}^{N} E_i. \tag{24}$$

The process of using a computer to generate particle histories can be performed in a way completely analogous to the actual physical process of particle transport through a medium. This direct simulation of the physical transport is called an *analog* Monte Carlo procedure. However, if the tally region is far from the source regions, most analog particle histories will make zero contribution to the tally, and thus a huge number of histories must be generated to obtain a statistically meaningful result. To reduce the number of histories, *nonanalog* Monte Carlo procedures can be used whereby certain biases are introduced in the generation of particle histories to increase the chances that a particle reaches the tally region. For example, source particles could be emitted preferentially toward the tally region instead of with the usual isotropic emission. Of course, when tallying such biased histories, corrections must be made to undo the bias so that a correct score is obtained. Many biasing schemes have been developed, and are generally called *variance* *reduction* methods since, by allowing more histories to score, the statistical uncertainty or variance in the average score is reduced.

The great advantage of the Monte Carlo approach, unlike discrete-ordinates, is that it can treat complex geometries. However, Monte Carlo calculations can be computationally extremely expensive, especially for deep-penetration problems. The stochastic contribution a single history makes to a particular score requires that a great many histories be simulated to achieve a good estimate of the expected or average score. If a tally region is many mean-free-path lengths from the source, very few histories reach the tally region and contribute to the score. Even with powerful variance reduction techniques, enormous numbers of histories often are required to obtain a meaningful score in deep-penetration problems.

Those readers interested in more comprehensive treatments of the Monte Carlo method will find rich resources. A number of monographs address Monte Carlo applications in radiation transport. Those designed for the specialists in nuclear-reactor computations are Goertzel and Kalos [67], Kalos [68], Kalos et al. [69], and Spanier and Gelbard [70]. More general treatments will be found in the books by Carter and Cashwell [71] Lux and Koblinger [72], and Dunn and Shultis [73]. Coupled photon and electron transport are addressed in the compilation edited by Jenkins et al. [74]. A very great deal of practical information can be gleaned from the manuals for Monte Carlo computer codes. Especially recommended are those for the EGS4 code [75], the TIGER series of codes [76], and the MCNP code [77].

Future Directions

In many respects, radiation shielding is a mature technological discipline. It is supported by a comprehensive body of literature and a diverse selection of computational resources. Indeed, the present availability of inexpensive computer clusters and the many sophisticated transport codes incorporating the most detailed physics models, modern data, and the ability to model complex geometries has reduced shielding practice in many cases to brute force calculation. Many shielding problems require such a computer approach; however, there are many routine shielding problems that can be effectively treated using the simplified techniques developed in the 1940s–1970s. Point-kernel methods are still widely used today. However, there are shielding problems for which no simplified approach is effective and transport methods must be employed. These include transmission of radiation through ducts and passages in structures, reflection of gamma rays from shielding walls and other structures, and transmission of beams of radiation obliquely incident on shielding slabs.

Despite the relative maturity of the discipline, one must not become complacent. There will continue to be advances in many areas. Undoubtedly, new computational resources will allow much more detailed 3-D graphical modeling of the shielding geometries and their incorporation into the transport codes. Likewise 3-D displays of output will allow much better interpretations of results. New capabilities will be added to Monte Carlo and discrete-ordinates codes. Hybrid codes employing both Monte Carlo and deterministic techniques will also be developed. More nuclear data will become available that will, for example, allow detailed analysis of actinides in spent fuel and correlation effects in nuclear data will allow better sensitivity analyses of results. Likewise, more information on material properties, especially in radiation resistance, will become known. Advances in microdosimetry will provide better understanding of cellular responses to single radiation particles and the effects of low-level radiation doses. A better understanding of radiation hormesis effects may lead to changes in radiation standards that will better reflect health effects of radiation. New sources of radiation in research and medicine will include energetic protons and neutrons. These developments require continuing attention and adoption into the radiation-shielding discipline.

Bibliography

Primary Literature

1. Shultis JK, Faw RE (2005) Radiation shielding technology. Health Phys 88:297–322
2. Mutscheller A (1925) Physical standards of protection against Roentgen-ray dangers. Am J Roentgenol Radiat Ther 13:65–70

3. National Council on Radiation Protection and Measurements (1941) Safe handling of radioactive luminous compounds. NBS handbook 27. NCRP report 5. US Government Printing Office, Washington, DC
4. Blizard EP, Abbott LS (eds) (1962) Reactor handbook, vol III, Part B, Shielding. Wiley, New York
5. Jaeger RH (ed) (1968) Engineering compendium on radiation shielding, vol 1, Shielding fundamentals and methods. Springer, New York
6. Haffner JW (1967) Radiation and shielding in space. Academic, New York
7. Shure K (1964) P-3 multigroup calculations of neutron attenuation. Nucl Sci Eng 19:310
8. National Council on Radiation Protection and Measurements (1971) Protection against neutron radiation. NCRP report 38. National Council on Radiation Protection and Measurements, Washington, DC
9. National Council on Radiation Protection and Measurements (1976) Structural shielding design and evaluation for medical use of x-rays and gamma rays of energies up to 10 MeV. NCRP report 49. National Council on Radiation Protection and Measurements, Washington, DC
10. Kocher DC (1981) Radioactive decay tables. Technical Information Center, U.S. Department of Energy; DOE/TIC 11026, Washington, DC
11. Weber DA, Eckerman KE, Dillman LT, Ryman JC (1989) MIRD: radionuclide data and decay schemes. Society of Nuclear Medicine, Medical Internal Radiation Dose Committee, New York
12. ANS (American Nuclear Society) (1991) American national standard gamma-ray attenuation coefficients and buildup factors for engineering materials. ANSI/ANS-6.4.3-1991. American Nuclear Society, La Grange Park
13. Caswell RS, Coyne JJ, Randolph ML (1980) Kerma factors for neutron energies below 30 MeV. Radiat Res 83: 217–254
14. International Commission on Radiological Protection (1987) Data for use in protection against external radiation. Publication 51. Annals of the ICRP 17(2/3). Pergamon, Oxford
15. International Commission on Radiological Protection (1996) Conversion coefficients for use in radiological protection against external radiation. Publication 74. Annals of the ICRP 26(3/4). Pergamon, Oxford
16. Adams ML, Larsen EW (2002) Fast iterative methods for discrete-ordinates particle transport calculations. Prog Nucl Energy 40(1):3–159
17. Shultis JK, Faw RE (2000) Radiation shielding. American Nuclear Society, La Grange Park
18. Rockwell T III (ed) (1956) Reactor shielding design manual. Van Nostrand, Princeton
19. Schaeffer NM (1973) Historical background. In: Schaeffer NM (ed) Reactor shielding. TID-25951. US Atomic Energy Commission, Washington, DC
20. Hirayama H (1987) Exposure buildup factors of high-energy gamma rays for water, concrete, iron, and lead. Nucl Technol 77:60–67
21. Faw RE, Shultis JK (1993) Absorbed dose buildup factors in air for 10–100 MeV photons. Nucl Sci Eng 114:76–80
22. Fano U, Spencer LV, Berger MJ (1959) Penetration and diffusion of x rays. In: Encyclopedia of physics, vol 38, Part 2. Springer, Berlin
23. Goldstein H (1959) Fundamental aspects of reactor shielding. Addison-Wesley, Reading
24. Spencer LV (1962) Structure shielding against fallout radiation from nuclear weapons. Monograph 42. National Bureau of Standards, Washington, DC
25. Takeuchi K, Tanaka S, Kinno M (1981) Transport calculation of gamma rays including bremsstrahlung by the discrete ordinates code PALLAS. Nucl Sci Eng 78:273–283
26. Takeuchi K, Tanaka S (1984) Buildup factors of gamma rays, including bremsstrahlung and annihilation radiation for water, concrete, iron, and lead. Nucl Sci Eng 87:478–489
27. Negin CA, Worku G (1998) Microshield v.5 user's manual. Grove Software, Lynchburg
28. Malenfant RE (1967) QAD: a series of point kernel general-purpose shielding programs. LA-3573. Los Alamos National Laboratory, Los Alamos
29. Price JH et al (1979) Utilization instructions for QADMOD-G. RRA-N7914. Radiation Research Association, Fort Worth. Available from RSICC as CCC565/QADMOD-GP
30. Litwin KA et al (1994) Improvements to the point kernel code QAD-CGGP: a code validation and user's manual. RC-1214 COG-94-65. AECL Research, Canada. Available from RSICC as CCC-645/QAD-CGGP-A
31. Malenfant RE (1990) G^3: a general purpose gamma-ray scattering code. LA-5176. Los Alamos National Laboratory, Los Alamos. Available from RSICC as CCC-564/G33-GP
32. Archer BR (1995) History of the shielding of diagnostic x-ray facilities. Health Phys 69:750–758
33. Simpkin DJ (1989) Shielding requirements for constant-potential diagnostic x-ray beams determined by a Monte Carlo calculation. Health Phys 56:151–154
34. Archer BR, Conway BJ, Quinn PW (1994) Attenuation properties of diagnostic x-ray shielding materials. Med Phys 21:1499–1507
35. Simpkin DJ (1995) Transmission data for shielding diagnostic x-ray facilities. Health Phys 68:704–709
36. Chilton AB (1971) Effect of material composition on neutron penetration of concrete slabs. Report 10425. National Bureau of Standards, Washington, DC
37. Roussin RW, Schmidt FAR (1971) Adjoint Sn calculations of coupled neutron and gamma-ray transport through concrete slabs. Nucl Eng Des 15:319–343
38. Roussin RW, Alsmiller RG Jr, Barish J (1973) Calculations of the transport of neutrons and secondary gamma rays through concrete for incident neutrons in the energy range 15 to 75 MeV. Nucl Eng Des 24:250–257

39. Wyckoff JM, Chilton AB (1973) Dose due to practical neutron energy distributions incident on concrete shielding slabs. Proceedings of 3rd international congress IRPA, American Nuclear Society, La Grange Park
40. Wang X, Faw RE (1995) Transmission of neutrons and secondary gamma rays through concrete slabs. Radiat Prot Dosim 60:212–222
41. Chilton AB, Huddleston CM (1963) A semi-empirical formula for differential dose albedo for gamma rays on concrete. Nucl Sci Eng 17:419–424
42. Chilton AB, Davisson CM, Beach LA (1965) Parameters for C-H albedo formula for gamma rays reflected from water, concrete, iron, and lead. Trans Am Nucl Soc 8:656
43. Chilton AB (1967) A modified formula for differential exposure albedo for gamma rays reflected from concrete. Nucl Sci Eng 27:481–482
44. Brockhoff RC (2003) Calculation of albedos for neutrons and photons. Ph.D. dissertation, Department of Mechanical and Nuclear Engineering, Kansas State University, Manhattan
45. Selph WE (1973) Albedos, ducts, and voids. In: Schaeffer NM (ed) Reactor shielding. TID-25951. US Atomic Energy Commission, Washington, DC
46. Simmons GL, Albert TE, Gritzner DM (1979) The SAI/EPRI information library. Report SAI-013-79-525-LJ. Science Applications Inc, La Jolla
47. Cain VR, Emmett MV (1979) BREESE-II: auxiliary routines for implementing the albedo option in the MORSE Monte Carlo code. ORNL/TM-6807. Oak Ridge National Laboratory, Oak Ridge
48. Maerker RE, Muckenthaler FJ (1966) Measurements and single-velocity calculations of differential angular thermal-neutron albedos for concrete. Nucl Sci Eng 26:339
49. Coleman WA, Maerker RE, Muckenthaler FJ, Stevens PJ (1967) Calculation of doubly differential current albedos for epicadmium neutrons incident on concrete and comparison of the subcadmium component with experiment. Nucl Sci Eng 27:411–422
50. Chandrasekhar S (1960) Radiative transfer. Dover, New York
51. Wells MB (1964) Reflection of thermal neutrons and neutron capture gamma rays from concrete. USAEC report RRA-M44. Radiation Research Associates, Fort Worth
52. Gomes IC, Stevens PN (1991) MORSE/STORM: a generalized albedo option for Monte Carlo calculations. ORNL/FEDC-91/1, TN. Oak Ridge National Laboratory, Oak Ridge
53. LeDoux JC, Chilton AB (1959) Gamma ray streaming through two-legged rectangular ducts. Nucl Sci Eng 11:362–368
54. National Council on Radiation Protection and Measurements (2003) Radiation protection for particle accelerator facilities. NCRP report 144. National Council on Radiation Protection and Measurements, Washington, DC
55. Lampley CM, Andrews MC, Wells MB (1988) The SKYSHINE-III procedure: calculation of the effects of structure design on neutron, primary gamma-ray, and secondary gamma-ray dose rates in air. RRA-T8209A. Radiation Research Associates, Fort Worth
56. Lampley CM (1979) The SKYSHINE-II procedure: calculation of the effects of structure design on neutron, primary gamma-ray, and secondary gamma-ray dose rates in air. RRA-T7901. Radiation Research Associates, Fort Worth
57. Brockhoff RC, Shultis JK, Faw RE (1996) Skyshine line-beam response functions for 20- to 100-MeV photons. Nucl Sci Eng 123:282–288
58. Gui AA, Shultis JK, Faw RE (1997) Response functions for neutron skyshine analysis. Nucl Sci Eng 125:111–127
59. Shultis JK, Faw RE, Bassett MS (1991) The integral line-beam method for gamma skyshine analysis. Nucl Sci Eng 107:228–245
60. Stedry MH (1994) A Monte Carlo line-beam calculation of gamma-ray skyshine for shielded sources. MS thesis, Kansas State University, Manhattan
61. Negin CA (1987) The microskyshine manual. Grove Software, Lynchburg
62. Carlson BG, Lathrop KD (1968) Transport theory, the method of discrete ordinates. In: Greenspan H, Kelber CN, Okrent D (eds) Computing methods in reactor physics. Gordon and Breach, New York
63. Duderstadt JJ, Martin WR (1979) Transport theory. Wiley, New York
64. Lewis EE, Miller WF (1984) Computational methods of neutron transport theory. Wiley, New York
65. Rhoades WA, Childs RL (1987) The TORT three-dimensional discrete ordinates neutron/photon transport code. ORNL-6268. Oak Ridge National Laboratory, Oak Ridge
66. Alcouffe RE, Baker RS, Dahl JA, Turner SA (2002) PARTISN user's guide. CCS-4, LA-UR-02-5633. Los Alamos National Laboratory, Transport Methods Group, Los Alamos
67. Goertzel G, Kalos MH (1958) Monte Carlo methods in transport problems. In: Progress in nuclear energy, ser 1, vol 2. Pergamon, New York
68. Kalos MH (1968) Monte Carlo integration of the adjoint gamma-ray transport equation. Nucl Sci Eng 33:284–290
69. Kalos MH, Nakache NR, Celnik JC (1968) Monte Carlo methods in reactor computations. In: Greenspan H, Kelber CN, Okrent D (eds) Computing methods in reactor physics. Gordon and Breach, New York
70. Spanier J, Gelbard EM (1969) Monte Carlo principles and neutron transport problems. Addison-Wesley, Reading
71. Carter LL, Cashwell ED (1975) Particle-transport simulation with the Monte Carlo method. TID-26607. Los Alamos National Laboratory, Los Alamos
72. Lux I, Koblinger LK (1991) Monte Carlo particle transport methods: neutron and photon calculations. CRC Press, Boca Raton
73. Dunn WL, Shultis JK (2011) Exploring the Monte Carlo method. Elsevier, Amsterdam
74. Jenkins TM, Nelson TM, Rindi A (1988) Monte Carlo transport of electrons and photons. Plenum, New York
75. Nelson WR, Hirayama H, Rogers DWO (1985) The EGS4 code system. SLAC-265. Stanford Linear Accelerator Center, Menlo Park

76. Halbleib JA, Kensek RP, Mehlhorn TA, Valdez GD, Seltzer SM, Berger MJ (1992) ITS Version 3.0: the integrated TIGER series of coupled electron/photon Monte Carlo transport codes. SAND91-1634. Sandia National Laboratories, Albuquerque

77. X-5 Monte Carlo Team (2003) MCNP—a general Monte Carlo n-particle transport code, version 5. LA-UR-03-1987 (vol 1: Overview and theory), LA-UR-03-0245 (vol 2: User's guide). Los Alamos National Laboratory, Los Alamos

Books and Reviews

Blizard EP, Abbott LS (eds) Reactor handbook, vol 3, Part B, Shielding. Wiley, New York

Faw RE, Shultis JK (1999) Radiological assessment: sources and doses. American Nuclear Society, La Grange Park

Goldstein H (1959) Fundamental aspects of reactor shielding. Addison-Wesley, Reading

Haffner JW (1967) Radiation and shielding in space. Academic, New York

ICRP (2007) The 2007 recommendations of the international commission on radiological protection. Report 103. Annals of the ICRP 37:2–4

ICRP (2008) Nuclear decay data for dosimetric calculations. Report 107. Annals of the ICRP 38(3):1–96

ICRU (1993) Quantities and units in radiation protection dosimetry. Report 51. International Commission on Radiation Units and Measurements, Bethesda

ICRU (1998) Fundamental quantities and units for ionizing radiation. Report 60. International Commission on Radiation Units and Measurements, Bethesda

Jaeger RG (ed) (1968–1975) Engineering compendium on radiation shielding. Shielding materials and design, vol 1; Shielding materials and designs, vol 2; Shield design and engineering, vol 3. Springer, New York

NCRP (2003) Radiation protection for particle accelerator facilities. Report 144. National Council on Radiation Protection and Measurements, Bethesda

NCRP (2005) Structural shielding design for medical x-ray imaging facilities. Report 147. National Council on Radiation Protection and Measurements, Bethesda

NCRP (2005) Structural shielding design and evaluation for megavoltage x- and gamma-ray radiotherapy facilities. Report 151. National Council on Radiation Protection and Measurements, Bethesda

Rockwell T III (ed) (1956) Reactor shielding design manual. Van Nostrand, Princeton

Schaeffer NM (ed) (1973) Reactor shielding. TID-25951, U.S. Atomic Energy Commission, Washington, DC

Shultis JK, Faw RE (2000) Radiation shielding. American Nuclear Society, La Grange Park

UN (1977, 1982, 1988, 1993, 2000) Reports of the United Nations Scientific Committee on the Effects of Atomic Radiation, New York. http://www.unscear.org/unscear/en/publications.html

Radiation Sources

Richard E. Faw[1], J. Kenneth Shultis[2]
[1]Department of Mechanical & Nuclear Engineering, Kansas State University, Winston Salem, NC, USA
[2]Department of Mechanical & Nuclear Engineering, Kansas State University, Manhattan, KS, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Radiation-Producing Reactions
Radioactivity
Sources of Neutrons
Sources of Gamma Photons
Sources of X-rays
Cosmic Rays, Solar Radiation, and Trapped Radiation Belts
Radiation Sources Used in Human Activities
Future Directions
Bibliography

Glossary

Absorbed dose A general term for the energy transferred from radiation to matter. Specifically, the absorbed dose is the amount of energy absorbed in a unit mass of matter from ionizing radiation. Units are the gray (Gy) and rad, respectively, equivalent to 1 J/kg and 100 ergs/g. Thus, 1 Gy equals 100 rad.

Activity The decay rate (expected number of nuclear transformations per unit time) in a radioactive sample. Units are the becquerel (Bq) equal to one decay per second, and the curie (Ci) equal to 3.7×10^{10} decays per second.

Alpha particle The nucleus of a ^{4}He atom, composed of two neutrons and two protons and denoted by α.

Committed dose The dose equivalent accumulated over the rest of a person's life following the ingestion or inhalation of radioactive material into the body.

Coulomb force The electrostatic force between two charges. It is proportional to the product of the charges and inversely proportional to the square of the distance between them. The force is attractive if the charges are of opposite sign, and repulsive if of like sign.

Dose equivalent A measure of the health risk associated with the absorption of radiation locally in the human body. It equals the local absorbed dose multiplied by a *quality factor* to correct for the relative biological effect associated with different radiations. Units are the sievert (Sv) or rem.

Effective dose An overall measure of the risk of cancer or hereditary illness associated with radiation exposure. It is a weighted average dose to multiple organs and tissues of the body, with weighting over both quality factor and the relative sensitivity of each organ or tissue. Units are the sievert (Sv) or rem for the dose in grays and rads, respectively.

Hadron A subatomic particle that reacts via strong nuclear forces. Hadrons include mesons (e.g., pions and kaons) and baryons (e.g., protons and neutrons). Hadrons do not include bosons (e.g., photons) and leptons (e.g., electrons, muons, and neutrinos).

Meson A subatomic particle, a subclass of hadrons, composed of an even number of other subatomic particles called quarks. Most important are the pi meson (pion) and K meson (kaon).

Nuclide A term used to refer to a particular atom or nucleus with a specific neutron number N and atomic (proton) number Z. The nuclide with N neutrons and Z protons and electrons is denoted as ${}^{A}_{Z}\mathrm{X}$ where X is the chemical symbol (determined by Z) and $A = Z + N$ is the mass number. If the nuclide is radioactive, it is called a *radionuclide*.

Photon A quantum of electromagnetic radiation with energy $E = hv$ where h is Planck's constant and v is the frequency. Photons produced by changes in the structure of a nucleus are called *gamma photons*, and those produced by atomic electron rearrangement are called *x-rays*.

Positron The antiparticle of the electron with the same mass m_e but with a positive charge equal in magnitude to the negative charge of the electron. A positron, denoted by β^+, quickly after its formation annihilates with an ambient electron converting the two electron masses into two photons each with energy $m_e c^2 = 0.511$ MeV.

Definition of the Subject

This entry treats sources of only ionizing radiation, such as electrons, protons, high-energy photons, neutrons, and similar radiations that have the ability to cause ionization, either directly or indirectly, and, thus, to induce chemical and physical changes along their passages through materials. Not included are sources of relatively lower frequency electromagnetic radiation from radio waves to ultraviolet light.

Characterization of sources requires characterization of radiations as well. Many sources encountered are radioactive isotopes of elements in the periodic table. These sources, as they decay radioactively, emit ionizing radiation. Some of the ionizing radiation is in the form of alpha particles, gamma rays, and x-rays, all characteristically monoenergetic in nature. Some is in the form of beta particles, distributed in energy but with well-defined maxima. Other sources are not directly associated with radioisotopes. X-ray machines and accelerators release ionizing radiation generally distributed in energy but with some monoenergetic components. The radiation belts surrounding the earth are composed of electrons and protons distributed in energy. Solar radiation and galactic cosmic rays are ionizing radiations widely distributed in type and energy. The nuclear fission process results in prompt emission of gamma rays and neutrons very widely distributed in energy. Fission also yields an extremely wide range of *fission products*, which radioactively decay over long periods of time.

Characterization of sources, to be meaningful, also begs discussion of radiation doses and radiation effects. There are acute effects accompanying high exposures. There are also known carcinogenic effects of human exposure and suspected mutational and hereditary effects. Therefore, a portion of this entry is devoted to examination of health effects associated with exposure to ionizing radiation.

This entry is divided into two parts. In the first several sections, the quantitative technical characterization of physical processes that produce ionizing radiation are reviewed. In latter sections, a qualitative examination is given of the various types of radiation sources encountered in the environment, workplace, laboratory, or medical facility.

Introduction

Throughout our lives, ionizing radiation is ever present, though rarely sensed. Radioactive sources are present in the food we eat, in the water we drink, and in the air we breathe. Most of these sources have but brief sojourns in our bodies, but some are taken up in bone and permanently retained. These sources, isotopes of elements in the periodic table, decay radioactively, emitting ionizing radiation most often in the form of gamma and x-rays, alpha particles, beta particles, and electrons. The ionization taking place in the body accounts for biological effects, good and bad.

Radiation also reaches us from sources outside our bodies. Radioactivity is present in our soils and minerals, and in our construction materials. Electromagnetic radiation of all wavelengths, including radio waves, microwaves, radar, and light, of both man-made and natural origins, constantly, bombard us. Photons are far more prevalent in number than atoms in our universe; for every nucleon there are about 10^9 photons. Cosmic rays and the subatomic debris they create during interactions in the atmosphere also impinge on us. Neutrinos from fusion reactions in our sun reach us in such numbers that tens of billions per second pass through every square centimeter of our skin. Most of this radiation, for example, neutrinos and radio waves, fortunately, passes harmlessly through us. Other radiation such as light and longer wavelength electromagnetic radiation usually interacts harmlessly with our tissues. However, shorter wavelength electromagnetic radiation, for example, ultraviolet light, x-rays and gamma rays, and charged particles produced by nuclear reactions can cause various degrees of damage to our cells.

The types and sources of radiation just described may be naturally occurring, may be a legacy of the era of nuclear weapons testing, or may be a result of human enterprise, for example, uranium in coal ash, radium in mine tailings, medical wastes, and fission or activation products in wastes from nuclear power production. All vary with latitude, longitude, and altitude – even on a small scale.

There are also population groups especially affected, but in different ways, by exposure to ionizing radiation from many sources. First among these in importance are patients and providers of medical radiation exposures. In the USA, as of 2006, collective effective doses to patients accruing from medical exposure, about 900,000 person-Sv annually, amount to almost half the total for the entire population. Computed tomography and nuclear medicine procedures dominate the exposure, with fluoroscopy and radiology accounting for about 230,000 person-Sv annual effective dose. Interventional fluoroscopy, while of great value to the patient, contributes in a major way to provider dose. Brachytherapy and beam therapy using photons, electrons, and protons lead to high therapeutic radiation exposures to patients, of course. Modern beam therapy utilizes exquisite beam shielding techniques to minimize doses to off-target patient tissues. Of all occupational groups, medical workers are greatest in number and accrue the highest collective doses, namely, 549 person-Sv annually in the USA. However, recordable individual worker annual doses are about 0.75 mSv as compared to 1.87 mSv for the fewer workers in commercial nuclear power.

Another population group consists of astronauts and aviation flight crews especially in high-altitude, international flights. Radiation from solar and galactic sources lead to high occupational radiation exposures. For astronauts, the life-threatening risk of solar-flare radiation exposures is a major concern.

Radiation sources used in industry affect another significant occupational group. Industrial radiography using x-ray, gamma-ray, or even neutron sources is an important component of occupational exposure. Other sources find wide use in measurement devices and sensor appliances.

For radiation to produce biological damage, it must first interact with tissue to alter molecular bonds and change the chemistry of the cells. Likewise, for radiation to produce damage in structural and electrical materials, it must cause interactions that disrupt crystalline and molecular bonds. Such radiation must be capable of creating ion–electron pairs and is termed *ionizing* radiation. Fast-moving charged particles, such as alpha particles, beta particles, and fission fragments, can directly ionize matter. Neutral particles, such as photons and neutrons, cannot interact directly with the electrons of the matter through they pass; rather they cause interactions that transfer some of their energy to charged secondary particles, which in turn produce ionization as they slow.

Radiation-Producing Reactions

Origins of Ionizing Radiation

Ionizing radiation is invariably the consequence of physical reactions, involving subatomic particles, at the atomic or nuclear level. The possible radiation-producing reactions are many, and usually, although not always, involve altering the configuration of neutrons and protons in an atomic nucleus or the rearrangement of atomic electrons about a nucleus. These reactions can be divided into two categories:

Radioactive decay. In the first type of radiation-producing reaction, the nucleus of an atom spontaneously changes its internal arrangement of neutrons and protons to achieve a more stable configuration. In such spontaneous *radioactive* transmutations, ionizing radiation is almost always emitted. The number of known different atoms, each with a distinct combination of Z and A is about 3,200. Of these, only 266 are stable and 65 are long-lived radioisotopes all of which are found in nature. The remaining nuclides have been made by humans and are radioactive with lifetimes much shorter than the age of the solar system. Both naturally occurring and manufactured radionuclides are the mostly commonly encountered sources of ionizing radiation.

Binary reactions. The second category of radiation-producing interactions involves two impinging atomic or subatomic particles that react to form one or more reaction products. Examples include neutrons interacting with nuclei of atoms, or photons interacting with nuclei or atomic electrons. Many binary reactions, in which an incident subatomic particle x strikes an atom or nucleus X, produce only two reaction products, typically a residual atom or nucleus Y and some subatomic particle y. These binary two-product reactions are often written as $X(x,y)Y$, for example, ${}^{14}_{7}\mathrm{N}(\alpha, p){}^{17}_{8}\mathrm{O}$.

Energetics of Radiation-Producing Reactions

In any nuclear reaction, total energy must be conserved. The total energy (kinetic plus rest-mass energy) of the initial particles must equal the total energy of the final products, that is, $\sum_i [E_i + m_i c^2] = \sum_i \left[E_i^{'} + m_i^{'} c^2\right]$, where E_i (E'_i) is the kinetic energy of the ith initial (final) particle with a rest mass m_i (m'_i), and c is the speed of light.

Any change in the total kinetic energy of particles before and after the reaction, ΔE, must be accompanied by an equivalent change in the total rest mass of the particles before and after the reaction, Δm, that are related by Einstein's famous equation $\Delta E = \Delta mc^2$. To quantify this change in the kinetic or rest-mass energies, a so-called Q-value is defined as

$$\begin{aligned} Q &= (\text{rest mass of initial particles})c^2 \\ &\quad - (\text{rest mass of final particles})c^2 \\ &= (\text{KE of final particles}) - (\text{KE of intial particles}) \end{aligned}$$

The Q value of a nuclear reaction may be either positive or negative. If the rest masses of the reactants exceed the rest masses of the products, the Q value of the reaction is positive with the decrease in rest mass being converted into a gain in kinetic energy. Such a reaction is *exoergic*. Radioactive decay is such a spontaneous exoergic nuclear reaction in which the Q-value energy is converted into the kinetic energy of the products.

Conversely, if Q is negative, the reaction is *endoergic*. For this case, kinetic energy of the initial particles is converted into rest-mass energy of the reaction products. The kinetic energy decrease equals the rest-mass energy increase. Such reactions cannot occur unless the colliding particles have at least a certain amount of kinetic energy, the so-called *threshold energy* for the reaction. For the binary, two-product reaction $X(x,y)Y$, the threshold kinetic energy of x incident on a stationary X is, neglecting Coulombic barrier effects, given approximately by

$$E_{th} \simeq -\left(1 + \frac{m_x}{m_X}\right)Q.$$

In any reaction, linear momentum must also be conserved. Thus, the momentum of the reaction products must equal that of the reactants. For two-product nuclear reactions, conservation of linear momentum requires that the products, depending on their recoil directions, have very definite amounts of kinetic energy. By contrast, for reactions with three or more products, there is no unique division of the reaction energy, and the products generally have a continuous distribution of kinetic energies.

Physical Characterization of Sources

The most fundamental type of source is a *point source*. Clearly, no real source can have zero size, but a real source can be approximated as a point source provided (1) that the volume is sufficiently small, that is, with dimensions much smaller than the dimensions of the attenuating medium between source and detector, and (2) that there is negligible interaction of radiation with the matter in the source volume. The second requirement may be relaxed if source characteristics are modified to account for source self-absorption and other source–particle interactions.

In general, a point source may be characterized as depending on energy, direction, and time. In almost all cases, time is not treated as an independent variable because the time delay between a change in the source and the resulting change in the radiation field is usually negligible. Therefore, the most general characterization of a point source used here is in terms of energy and direction. Most radiation sources treated in shielding practice are isotropic, so that source characterization requires only knowledge of energy dependence. Radioisotope sources are certainly isotropic, as are fission sources and capture gamma-ray sources.

A careful distinction must be made between the activity of a radioisotope and its source strength. *Activity* is precisely defined as the expected number of atoms undergoing radioactive transformation per unit time. It is *not* defined as the number of particles emitted per unit time. Decay of two very common laboratory radioisotopes illustrate this point. Each transformation of ^{60}Co, for example, results in the emission of two gamma rays, one at 1.173 MeV and the other at 1.333 MeV. Each transformation of ^{137}Cs, accompanied by a transformation of its decay product $^{137\mathrm{m}}$Ba, results in emission of a 0.662-MeV gamma ray with probability 0.85.

The SI unit of activity is the becquerel (Bq), equivalent to one transformation per second. In medical and health physics, radiation source strengths are commonly calculated on the basis of *accumulated activity*, Bq s. Such time-integrated activities account for the cumulative number of transformations in some biological entity during the transient presence of radionuclides in the entity. Of interest in such circumstances is not the time-dependent dose rate to that entity or some other nearby region, but rather the total dose accumulated during the transient. Similar practices are followed in dose evaluation for reactor transients, solar flares, nuclear weapons, and so on.

Radiation sources may be distributed along a line, over an area, or within a volume. Source characterization requires, in general, spatial and energy dependence. Occasionally, it is necessary to include angular dependence. This is especially true for effective area sources associated with computed angular flows across certain planes. Energy dependence may be discrete, such as for radionuclide sources, or continuous, as for bremsstrahlung or fission neutrons and photons.

Radioactivity

Radioactive Decay Dynamics

The decay of a radioactive nuclide is a stochastic phenomenon. The time an individual radionuclide decays cannot be predicted; rather, only the probability of decay in a specified time interval can be predicted. The rate at which a sample of a large number of identical radionuclides decays is determined by the *radioactive decay constant* λ for the nuclide. This constant is the probability, per unit time, that a radionuclide decays in an infinitesimal time interval. That λ is constant for a given radionuclide species implies that the expected number of radionuclides, $N(t)$, at time t is $N(t) = N(0)e^{-\lambda t}$, where $N(0)$ is the initial number of radionuclides in the sample. The exponential decay of radionuclides is sometimes called the *radioactive decay law*.

Generally, the number of radionuclides in a sample is not of interest. Rather the *activity* $A(t)$ or rate at which a radionuclide sample decays, $\mathrm{d}N(t)/\mathrm{d}t$, is desired since this quantity determines the rate of radiation emission from the sample. From the radioactive decay law, it is found that $\mathrm{d}N(t)/\mathrm{d}t = \lambda N(t) \equiv A(t)$, so that the activity of a radionuclide sample also decays exponentially, that is, $A(t) = A(0)e^{-\lambda t}$. The standard unit of activity is the becquerel (Bq) equal to one radioactive decay per second. The traditional unit is the curie (Ci) = 3.7×10^{10} Bq (approximately the activity of 1 g of ^{226}Ra).

The rate at which a radioactive sample decays is commonly described by its *half-life* $T_{1/2}$. The half-life is the time required for half of the sample to decay, or, equivalently, for the sample activity to halve. From the radioactive decay law, it is found $T_{1/2} = \ln 2/\lambda \simeq 0.693/\lambda$.

Types of Radioactive Decay

There are several types of spontaneous changes (or *transmutations*) that can occur in radioactive nuclides. In each transmutation, the nucleus of the parent atom ${}^{A}_{Z}\mathrm{P}$ is altered in some manner and one or more particles of radiation are emitted. If the number of protons in the nucleus is changed, then the number of orbital electrons in the daughter atom D must subsequently also be changed, either by releasing an electron to or absorbing an electron from the ambient medium. The most commonly encountered types of radioactive decay are

Gamma decay (γ) ${}^{A}_{Z}\mathrm{P}^* \rightarrow{}^{A}_{Z}\ \mathrm{P} + \gamma$: An excited nucleus decays (usually within 10^{-8} s) to its ground state by the emission of one or more gamma photons. The excited parent is often the product of radioactive decay or a binary nuclear reaction.

Isomeric transition (IT) ${}^{Am}_{Z}\mathrm{P}^* \rightarrow {}^{A}_{Z}\mathrm{P} + \gamma$: This is a special case of gamma decay, in which the excited parent has a lifetime much greater than usual nuclear lifetimes (10^{-8} s), ranging from seconds to thousands of years. Such a long-lived excited nucleus is said to be *metastable* and is called an *isomer*.

Internal conversion (IC) ${}^{A}_{Z}\mathrm{P}^* \rightarrow [{}^{A}_{Z}\mathrm{P}]^{+1} + \mathrm{e}^-$: The excitation energy of a nucleus is used to eject an orbital electron (usually a *K*-shell electron).

Alpha decay (α) ${}^{A}_{Z}\mathrm{P} \rightarrow {}^{Z-4}_{Z-2}\mathrm{D} + \alpha$: An α particle is emitted leaving the daughter with 2 fewer neutrons and 2 fewer protons than the parent. The transition often is to an excited nuclear state of the daughter which decays by emission of one or more gamma photons.

Beta decay (β^-) ${}^{A}_{Z}\mathrm{P} \rightarrow {}^{A}_{Z+1}\mathrm{D} + \beta^- + \bar{v}$: In effect, a neutron in the nucleus decays to a proton. An electron (β^-) and antineutrino ($\bar{v}$) are emitted, which share the decay energy. The daughter is often produced in an excited nuclear state and subsequently emits gamma photons.

Positron decay (β^+) ${}^{A}_{Z}\mathrm{P} \rightarrow {}^{A}_{Z-1}\mathrm{D} + \beta^+ + v$: In effect, a proton in the nucleus changes into a neutron. A positron (β^+) and neutrino (v) are emitted, which share the decay energy. If the daughter is produced in an excited state, gamma decay results. The emitted positron, after slowing in the ambient medium, annihilates with an ambient electron producing two 0.511-MeV gamma rays.

Electron capture (EC) ${}^{A}_{Z}\mathrm{P} \rightarrow {}^{A}_{Z-1}\mathrm{D}^* + v$: An orbital electron is absorbed by the nucleus, converts a nuclear proton into a neutron, emits a neutrino (v), and, generally, leaves the nucleus in an excited state, which decays by the emission of one or more gamma photons.

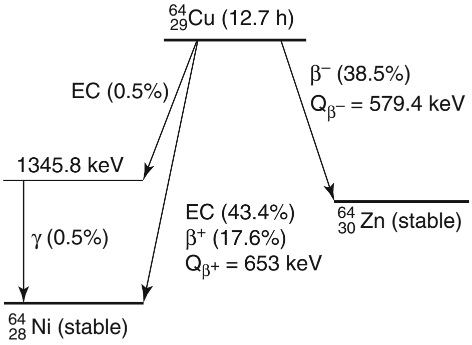
Spontaneous fission (SP) ${}^{A}_{Z}\mathrm{P} \rightarrow {}^{A_H}_{Z_H}\mathrm{D}_H + {}^{A_L}_{Z_L}\mathrm{D}_L + n({}^{1}_{0}\mathrm{n}) + m(\gamma)$: A heavy nucleus spontaneously splits or fissions into a heavy (*H*) and light (*L*) fission fragment. The fission fragments are produced in highly excited nuclear states and decay by prompt neutron and gamma photon emission within 10^{-13} s of the fission event, releasing, on the average, *n* neutrons and *m* γ photons. The resulting *fission products* are usually radioactive and undergo a chain of β^- decays releasing several delayed gamma photons and beta particles until a stable nucleus is reached. In some instances, ternary rather than binary fission takes place, releasing a light product such as tritium.

Many radionuclides decay by more than a single decay mechanism. For example, electron capture is always in competition with positron decay. An example of a radionuclide that decays by three mechanisms is ^{64}Cu whose decay scheme is shown in Fig. 1.

In any radioactive decay that alters the proton number *Z*, electron rearrangements necessarily result. The resulting cascade of orbital electrons to lower energy levels results in emission of x-rays and, in competition, ejection of what are called *Auger electrons*.

Naturally Occurring Radionuclides

Singly Occurring Primordial Radionuclides Of the many radioactive species present when the earth was formed, only those very few with half-lives comparable to the age of the earth remain in the environment. Of these few primordial radionuclides not belonging to a decay chain, only ^{40}K and ^{87}Rb contribute



Radiation Sources. Figure 1
The radioactive decay scheme for ^{64}Cu per decay, on average a beta particle of maximum energy 0.579 MeV is emitted with a probability of 0.385, a positron of maximum energy 0.653 MeV is emitted with a probability of 0.176, and a gamma ray of energy 1.346 MeV is emitted with a probability of 0.005. Source: NUDAT 2.5, National Nuclear Data Center, Brookhaven National Laboratory

significantly to human exposure. Of minor consequence are the nuclides ^{138}La, ^{147}Sm, and ^{176}Lu. The radionuclide ^{87}Rb has a half-life of 4.8 × 10^{10} years and decays by beta-particle emission. In the human body, its main impact is on bone-surface cells. The radionuclide ^{40}K is a major contributor to human exposure from natural radiation. Present in an isotopic abundance of 0.0118%, it has a half-life of 1.227 × 10^9 years, decaying both by electron capture and beta-particle emission. Annual human doses are about 140 μGy to bone surface, 170 μGy on average to soft tissue, and 270 μGy to red marrow [1]. ^{40}K also contributes in a major way to external exposure. The population-average specific activity of the nuclide in soil is 420 Bq/kg [2]. Based on a soil density of 1,600 kg/m^3, and dose conversion factors from [3], ^{40}K in the soil contributes 120 μSv effective dose annually.

Decay Series of Terrestrial Origin Two actinide decay series, identified by the long-lived parents ^{238}U and ^{232}Th contribute appreciably to human exposure to natural radiation. Another series headed by ^{235}U contributes very little. Members of the two important series are listed in Table 1. Many of the radionuclides in these series decay by emission of alpha particles with energies from 4 to 6 MeV. Others in the series emit beta particles accompanied by gamma rays. With long-lived parent radionuclides and short-lived daughter products, the chain might be thought to exist in a state of secular equilibrium, that is, each component having the same decay rate per unit volume of the host medium. However, some of the chain members are more soluble than others, and some are gaseous. Thus, unless the host medium is a rigid solid, such as granite, decay rates are far from equilibrium state.

Ingestion of elements in the uranium and thorium decay chains is unavoidable. Table 2 illustrates the consequences in terms of the committed effective dose incurred by ingestion but perhaps experienced long thereafter.

The portions of the series headed by the gases ^{220}Rn and ^{222}Rn are of special importance in public health. The gases escape from soil and rock into the atmosphere and into the airspace within homes. Their daughter products, some of which emit alpha particles, may be inhaled, with risk of radiation damage to radiation-sensitive cells in the lungs potentially leading to lung cancer. ^{222}Rn and its daughters ordinarily present a greater hazard than ^{220}Rn (thoron) and its daughters, largely because the much shorter half-life of ^{220}Rn makes decay more likely prior to release into the atmosphere. Globally, the mean annual effective dose equivalent due to ^{222}Rn daughters is about 1 mSv (100 mrem) while that due to ^{220}Rn daughters is estimated to be about 0.2 mSv (20 mrem).

Cosmogenic Radionuclides Cosmic-ray interactions with constituents of the atmosphere, sea, or earth, but mostly with the atmosphere, lead directly to radioactive products. Capture of secondary neutrons produced in primary interactions of cosmic rays, leads to the formation of many more radionuclides. Of the nuclides produced in the atmosphere, only ^{3}H, ^{7}Be, ^{14}C, and ^{22}Na contribute appreciably to human radiation exposure.

Over the past century, combustion of fossil fuels and the emission of carbon dioxide not containing ^{14}C has diluted the cosmogenic content of ^{14}C in the environment. Moreover, since World War II, artificial introduction of ^{14}C, ^{3}H, and other nuclides into the environment by human activity has been significant, especially as a result of atmospheric nuclear tests. Consequently, these radionuclides no longer exist in natural equilibria in the environment.

Radiation Sources. Table 1 Principal radioisotopes in two naturally occurring primordial decay series. Source: NUDAT 2.5, National Nuclear Data Center, Brookhaven National Laboratory

Thorium series			Uranium series		
Nuclide and decay mode		Half-life[a] $T_{1/2}$	Nuclide and decay mode		Half-life[a] $T_{1/2}$
$\mathbf{{}^{232}_{90}Th}$	α	**14.05 Gy**	$\mathbf{{}^{238}_{92}U}$	α	**4.468 Gy**
${}^{228}_{88}\mathrm{Ra}$	β	5.75 years	${}^{234}_{90}\mathrm{Th}$	β	24.10 days
${}^{228}_{89}\mathrm{Ac}$	β	6.15 h	${}^{234m}_{91}\mathrm{Pa}$	β	1.159 min
${}^{228}_{90}\mathrm{Th}$	α	1.912 years	${}^{234}_{92}\mathrm{U}$	α	245.5 ky
${}^{224}_{88}\mathrm{Ra}$	α	3.66 days	${}^{230}_{90}\mathrm{Th}$	α	75.4 ky
${}^{220}_{86}\mathrm{Rn}$	α	55.6 s	${}^{226}_{88}\mathrm{Ra}$	α	1,600 years
${}^{216}_{84}\mathrm{Po}$	α	0.145 s	${}^{222}_{86}\mathrm{Rn}$	α	3.8235 days
${}^{212}_{82}\mathrm{Pb}$	β	10.64 h	${}^{218}_{84}\mathrm{Po}$	α, β	3.098 min
${}^{212}_{83}\mathrm{Bi}$	α, β	60.55 min	${}^{214}_{82}\mathrm{Pb}$	β	26.8 min
${}^{212}_{84}\mathrm{Po}$	α	0.299 μs	${}^{214}_{83}\mathrm{Bi}$	α, β	19.9 min
${}^{208}_{81}\mathrm{Ti}$	β	3.053 min	${}^{214}_{84}\mathrm{Po}$	α, β	164 μs
${}^{208}_{82}\mathrm{Pb}$		∞	${}^{210}_{82}\mathrm{Pb}$	β	22.2 years
			${}^{210}_{83}\mathrm{Bi}$	α, β	5.012 days
			${}^{210}_{84}\mathrm{Po}$	α	138.4 days
			${}^{206}_{82}\mathrm{Pb}$		∞

[a] Gy = 10^9 years, ky = 10^3 years, μs = 10^{-6} s

Radiation Sources. Table 2 Annual intake and effective dose from ingestion of uranium, thorium, and daughters. For population weighted averages, apply 5% infants, 30% children, and 65% adults

	Activity intake (Bq/year)			Committed effective dose (μSv/year)		
Nuclide	Infants	Children	Adults	Infants	Children	Adults
^{238}U	1.9	3.8	5.7	0.23	0.26	0.25
^{234}U	1.9	3.8	5.7	0.25	0.28	0.28
^{230}Th	1.0	2.0	3.0	0.42	0.48	0.58
^{226}Ra	7.8	15	22	7.5	12	8.0
^{210}Pb	11	21	30	40	40	28
^{210}Po	21	39	58	180	100	85
^{232}Th	0.6	1.1	1.7	0.26	0.32	0.36
^{228}Ra	5.5	10	15	31	40	21
^{228}Th	1.0	2.0	3.0	0.38	0.30	0.25
^{235}U	0.1	0.2	0.2	0.011	0.011	0.011
Total				260	200	110

Source: Reference [2].

The tritium ^{3}H nuclide is produced mainly from interactions of neutrons with nitrogen and oxygen. Tritium has a half-life of 12.3 years and, upon decay, releases one low-energy beta particle of mean energy 5.7 keV. Tritium exists in nature almost exclusively in water form (HTO) but, in foods, may be partially incorporated into organic compounds. The nuclide ^{14}C is produced mainly from the interactions of neutrons with nitrogen in the atmosphere. It exists in the atmosphere as CO_2, but the main reservoir is the ocean. It has a half-life of 5,700 years and decays by beta particle emission of mean energy 49.5 keV.

^{7}Be is also produced by cosmic ray interactions with nitrogen and oxygen in the atmosphere. It decays by electron capture, 10.4% of which events yield a 0.478-MeV gamma ray. ^{22}Na decays by positron emission (90%) and electron capture (10%). The positron emission yields two annihilation photons as well as a positron of mean energy 215 keV. Both the positron emission and electron capture yield a 1.275-MeV gamma ray. Table 3 lists natural inventories, atmospheric concentrations, and effective doses to populations. Note that 1 PBq = 10^{15} Bq.

Sources of Neutrons

Fission Neutrons

Many heavy nuclides fission after the absorption of a neutron, or even spontaneously, producing several energetic fission neutrons. Almost all of the fast neutrons produced from a fission event are emitted within 10^{-14} s of the fission event, and are called *prompt neutrons*. Generally, less than 1% of the total fission neutrons are emitted as *delayed neutrons*, which are produced by the neutron decay of fission products at times up to many seconds or even minutes after the fission event. As the energy of the neutron which induces the fission in a heavy nucleus increases, the average number of fission neutrons also increases. For example, the fission of ^{235}U by a thermal neutron (average energy 0.025 eV) produces, on the average, 2.43 fission neutrons. A fission caused by a 10-MeV neutron, by contrast, yields 3.8 fission neutrons. For ^{239}Pu, fission by thermal or 10 MeV neutrons yield 2.87 or 4.2 neutrons. The fission of ^{238}U is induced only by fast neutrons, a 10-MeV neutron yielding 3.9 fission neutrons.

Radiation Sources. Table 3 Cosmogenic radionuclides and consequential mean doses to the population

Nuclide	Half-life	Global inventory (PBq)	Troposphere conc. (mBq/m^3)	Annual effective dose (μSv/year)
^{3}H	12.32 years	1,275	1.4	0.01
^{7}Be	53.22 days	413	12.5	0.03
^{14}C	5,700 years	12,750	56.3	12
^{22}Na	2.602 years	0.44	0.0021	0.15

Source: Reference [2].

Since the advent of fission reactors, many transuranic isotopes have been produced in significant quantities. Many of these isotopes have appreciable spontaneous fission probabilities, and consequently they can be used as very compact sources of fission neutrons. For example, 1 g of ^{252}Cf releases 2.3 × 10^{12} neutrons per second, and very intense neutron sources can be made from this isotope, limited in size only by the need to remove the fission heat through the necessary encapsulation. Almost all spontaneously fissioning isotopes decay much more frequently by α emission than by fission.

The energy dependence of the fission neutron spectrum has been investigated extensively, particularly for the important isotope ^{235}U. All fissionable nuclides produce prompt-fission neutrons with energy frequency distributions that go to zero at low and high energies, reaching a maximum at about 0.7 MeV, and have an average energy of about 2 MeV. The fraction of prompt fission neutrons emitted per unit energy about E, $\chi(E)$, can be described quite accurately by a Watt distribution

$$\chi(E) = ae^{-E/b}\sinh\sqrt{cE},$$

where the parameters a, b, and c depend on the fissioning isotope. For example, a = 0.5535 MeV, b = 1.0347 MeV, and c = 1.6214 MeV^{-1} for thermal-neutron fission of ^{235}U, whose fission-neutron spectrum is often used as an approximation for other fissioning isotopes.

Fusion Neutrons

Neutrons can be produced as products of nuclear reactions in which energetic charged particles hit target atoms. Most such reactions require accelerators to produce the energetic charged particles and, hence, such neutrons are to be encountered only near accelerator targets.

One major exception to the insignificance of charged-particle-induced reactions are those in which light elements fuse exoergically to yield a heavier nucleus and which are accompanied quite often by the release of energetic neutrons. The resulting fusion neutrons are usually the major source of radiation to be shielded against. The two neutron-producing fusion reactions of most interest in the development of thermonuclear fusion power are

$$^{2}\mathrm{H} + {}^{2}\mathrm{H} \rightarrow {}^{3}\mathrm{He}(0.82\ \mathrm{MeV}) + {}^{1}n(2.45\ \mathrm{MeV})$$

$$^{2}\mathrm{H} + {}^{3}\mathrm{H} \rightarrow {}^{4}\mathrm{He}(3.5\ \mathrm{MeV}) + {}^{1}n(14.1\ \mathrm{MeV}).$$

When these reactions are produced by accelerating one nuclide toward the other, the velocity of the center of mass must first be added to the center-of-mass neutron velocity before determining the neutron energy in the laboratory coordinate system. In most designs for fusion power, the velocity of the center of mass is negligible, and the concern is with monoenergetic 2.45- or 14.1-MeV fusion neutrons. The 14.1-MeV fusion neutrons are also produced copiously in a thermonuclear explosion.

A beam of relatively low-energy deuterons (100–300 keV) incident on a deuterium or tritium target can produce a significant number of thermonuclear neutrons. Thus, these D–D or D–T reactions are used in relatively compact accelerators, called *neutron generators*, in which deuterium ions are accelerated through a high voltage (100–300 kV) and allowed to fall on a thick deuterium- or tritium-bearing target. Typically, in such devices, a 1-mA beam current produces up to 10^{9} 14-MeV neutrons per second from a thick tritium target.

Photoneutrons

A gamma photon with energy sufficiently large to overcome the neutron binding energy (about 7 MeV in most nuclides) may cause a (γ,n) reaction. Very intense and energetic photoneutron production can be realized in an electron accelerator where the bombardment of an appropriate target material with the energetic electrons produces intense bremsstrahlung (see “Sources of X-rays”) with a distribution of energies up to that of the incident electrons. The probability a photon will cause a (γ,n) reaction increases with the photon energy, reaching a maximum over a broad energy range of approximately 20–23 MeV for light nuclei ($A \lesssim 40$) and 13–18 MeV for medium and heavy nuclei. The peak energy of this broad (often called *giant*) nuclear resonance can be approximated by 80 $A^{-1/3}$ MeV for $A > 40$. The width of the resonance varies from about 10 MeV for light nuclei to 3 MeV for heavy nuclei. Consequently, in medical or accelerator facilities that produce photons with energies above about 15 MeV, neutron production in the surrounding walls can lead to a significant neutron field.

However, the gamma photons produced in radioactive decay of fission and activation products in nuclear reactors generally have energies too low, and most materials have a photoneutron threshold too high for photoneutrons to be of concern. Only for the light elements ^{2}H, ^{6}Li, ^{7}Li, ^{9}Be, and ^{12}C are the thresholds for photoneutron production sufficiently low that these secondary neutrons may have to be considered. In heavy-water- or beryllium-moderated reactors, the photoneutron source may be very appreciable, and the neutron field deep within an hydrogenous shield is often determined by photoneutron production in deuterium, which constitutes about 0.015 atom percent of the hydrogen. Capture gamma photons arising from neutron absorption have particularly high energies, and thus may also cause a significant production of energetic photoneutrons.

The photoneutron mechanism can be used to create laboratory neutron sources by mixing intimately a beryllium or deuterium compound with a radioisotope that decays with the emission of high-energy photons. Alternatively, the encapsulated radioisotope may be surrounded by a beryllium- or deuterium-bearing shell. A common reactor photoneutron source is an antimony–beryllium mixture, which has the advantage of being rejuvenated by exposing the source to the neutrons in the reactor to transmute the stable ^{123}Sb into the required ^{124}Sb isotope (half-life of 60.2 days).

One very attractive feature of such (γ,n) sources is the nearly monoenergetic nature of the neutrons if the photons are monoenergetic. However, in large sources, the neutrons may undergo significant scattering in the source material and thereby degrade the nearly monoenergetic nature of their spectrum. These photoneutron sources generally require careful use because of their inherently large photon emission rates. Because nominally only one in a million high-energy photons actually interacts with the source material to produce a neutron, these sources generate gamma rays that are of far greater biological concern than are the neutrons.

Alpha-Neutron Sources

Many compact laboratory neutron sources use energetic alpha particles from various radioisotopes (*emitters*) to induce (α,n) reactions in appropriate materials (*converters*). Although a large number of nuclides emit neutrons if bombarded with alpha particles of sufficient energy, the energies of the alpha particles from radioisotopes are capable of penetrating the potential barriers of only the lighter nuclei.

Of particular interest are those light isotopes for which the (α,n) reaction is exoergic ($Q > 0$) or, at least, has a low threshold energy. For endoergic reactions ($Q > 0$), the threshold alpha energy is $-\ Q\ (1 + 4/A)$. Thus, for an (α,n) reaction to occur, the alpha particle must (1) have enough energy to overcome the repulsive Coulombic force field of the nucleus, and (2) exceed the threshold energy for the reaction. Converter materials used to make practical (α,n) sources include lithium, beryllium, boron, carbon, fluorine, and sodium.

The converter nuclides ^{18}O and ^{19}F are responsible for neutron production in many areas of the nuclear fuel cycle. Alpha particles emitted by uranium and plutonium range between 4 and 6 MeV in energy and can cause (α,n) neutron production when in the presence of oxygen or fluorine. In particular, (α,n) neutrons often dominate the spontaneous fission neutrons in UF_6 or in aqueous mixtures of uranium and plutonium such as found in nuclear waste.

A neutron source can be fabricated by mixing intimately a light converter element, such as lithium or beryllium, with a radioisotope that emits energetic alpha particles. Most of the practical alpha emitters are actinide elements, which form intermetallic compounds with beryllium. Such a compound, for example, $PuBe_{13}$, ensures both that the emitted alpha particles immediately encounter converter nuclei, thereby producing a maximum neutron yield, and that the radioactive actinides are bound into the source material, thereby reducing the risk of leakage of the alpha-emitting component.

The neutron yield from an (α,n) source varies strongly with the converter material, the energy of the alpha particle, and the relative concentrations of the emitter and converter elements. The degree of mixing between converter and emitter and the size, geometry, and source encapsulation may also affect the neutron yield. For example, a ^{239}Pu/Be source has an optimum neutron yield of about 60 neutrons per 10^6 primary alpha particles.

The energy distributions of neutrons emitted from (α,n) sources are continuous below some maximum neutron energy with definite structure at well-defined energies determined by the energy levels of the converter and the excited product nuclei. The use of the same converter material with different alpha emitters produces similar neutron spectra with different portions of the same basic spectrum accentuated or reduced as a result of the different alpha-particle energies. Average energies of neutrons typically are several MeV. For example, the neutrons produced by a ^{239}Pu/Be source have an average energy of 4.6 MeV.

Activation Neutrons

A few highly unstable nuclides decay by the emission of a neutron. The delayed neutrons associated with fission arise from such decay of the fission products. However, there are nuclides other than those in the fission-product decay chain which also decay by neutron emission. Only one of these nuclides, ^{17}N, is of importance in nuclear reactor situations. This isotope is produced in water-moderated reactors by an (*n,p*) reaction with ^{17}O (threshold energy, 8 MeV). The decay of ^{17}N by beta emission (half-life 4.4 s) produces ^{17}O in a highly excited state, which in turn decays rapidly by neutron emission. Most of the decay neutrons are emitted within ±0.2 MeV of the most probable energy of about 1 MeV, although neutrons with energies up to 2 MeV may be produced.

Spallation Neutron Sources

In a spallation neutron source, pulses of very energetic protons (up to 1 GeV), produced by an accelerator, strike a heavy metal target such as mercury or liquid bismuth. Such an energetic proton when it strikes a target nucleus "spalls" or knocks out neutrons. Additional neutrons boil off as the struck nucleus heats up. Typically, 20–30 neutrons are produced per spallation reaction. These pulses of neutrons are then slowed down or thermalized by passing them through cells filled with water, or even liquid hydrogen if very slow neutrons are needed.

Sources of Gamma Photons

Radioactive Decay

Radioactive sources serve a wide variety of purposes in educational, medical, research, industrial, governmental, and commercial activities. The radionuclides in these sources almost always leave their decay daughters in excited nuclear states whose subsequent transitions to lower-energy states usually result in the emission of one or more gamma photons.

Prompt Fission Photons

The fission process produces copious gamma photons either within the first 6×10^{-8} s after the fission event (the *prompt fission gamma photons*) or from the subsequent decay of the fission products. These photons are of extreme importance in the shielding and gamma-heating calculations for a nuclear reactor. Consequently, much effort has been directed toward determining their nature.

Most investigations of prompt fission gamma photons have centered on the thermal-neutron-induced fission of ^{235}U. For this nuclide, it has been found that the number of prompt fission photons is 8.13 ± 0.35 photons per fission over the energy range 0.1–10.5 MeV, and the energy carried by this number of photons is 7.25 ± 0.26 MeV per fission. The energy spectrum of prompt gamma photons from the thermal fission of ^{235}U between 0.1 and 0.6 MeV is approximately constant at 6.6 photons MeV^{-1} fission^{-1}. At higher energies, the spectrum falls off sharply with increasing energy. The measured energy distribution of the prompt fission photons can be represented by the following empirical fit over the range 0.1–10.5 MeV:

$$N(E) = \begin{cases} 6.6 & 0.1 < E < 0.6 \text{ MeV} \\ 20.2e^{-1.78E} & 0.6 < E < 1.5 \text{ MeV} \\ 7.2e^{-1.09E} & 1.5 < E < 10.5 \text{ MeV}, \end{cases}$$

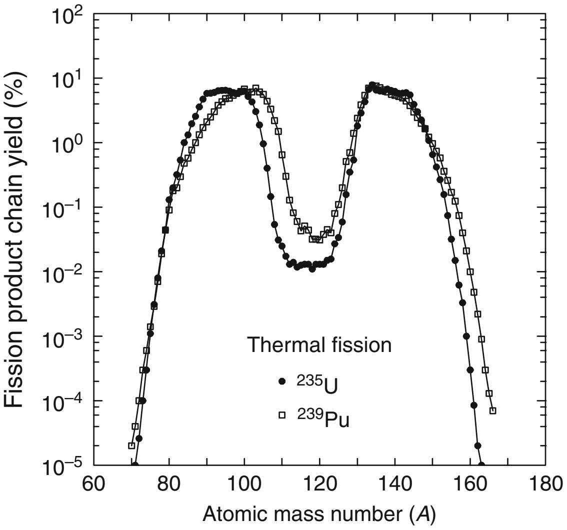
where E is in MeV and $N(E)$ is in units of photons MeV^{-1} fission^{-1}.

Investigation of ^{233}U, ^{239}Pu, and ^{252}Cf indicates that the prompt fission photon energy spectra for these isotopes resembles very closely that for ^{235}U, and hence for most purposes, it is reasonable to use the ^{235}U spectrum for other fissioning isotopes.

Fission-Product Photons

With the widespread application of nuclear fission, an important concern is the consideration of the very long lasting gamma activity produced by the decay of fission products.

In the fission process, most often two fragments are produced (*binary fission*) with a distribution in mass shown in Fig. 2. About 0.3% of the time, a third light fragment is produced (*ternary fission*), most often ^{3}H. As seen in Fig. 2, the mass distribution or *fission-product chain yield* is bimodal, with many products



Radiation Sources. Figure 2
The probability (%) that a fission product with mass number *A* is produced in the thermal-neutron-induced fission of ^{235}U and ^{239}Pu

having atomic mass number around 95 and around 140. Among the former are the important long-lived radionuclide ^{90}Sr, several isotopes of the halogen bromine, and various isotopes of the noble gas krypton. Among the heavy fragments are the important long-lived radionuclide ^{137}Cs, radioisotopes of halogen iodine, notably ^{131}I, and isotopes of the noble gas xenon. The fission-products are neutron-rich and decay almost exclusively by β^- emission, often forming long decay chains. From the range of mass numbers produced (see Fig. 2), about 100 different decay chains are formed. An example of a short chain is

$$^{140}_{54}\mathrm{Xe}\xrightarrow[16\,\mathrm{s}]{\beta^-}{}^{140}_{55}\mathrm{Cs}\xrightarrow[66\,\mathrm{s}]{\beta^-}{}^{140}_{56}\mathrm{Ba}\xrightarrow[12.8\,\mathrm{days}]{\beta^-}$$

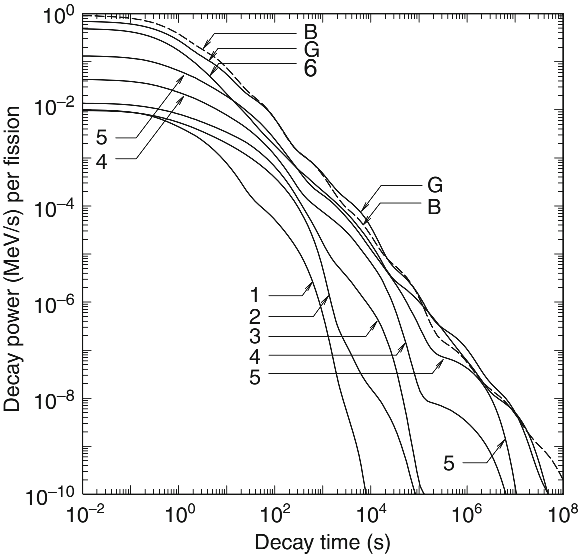
$$^{140}_{57}\mathrm{La}\xrightarrow[40\,\mathrm{h}]{\beta^-}{}^{140}_{58}\mathrm{Ce(stable)}.$$

The total gamma-ray energy released by the fission product chains is comparable to that released as prompt fission gamma photons. The gamma-ray energy release rate declines rapidly in the time after fission. About three-fourths of the delayed gamma-ray energy is released in the first 1,000 s after fission. In most calculations involving spent nuclear fuel, the gamma activity at several months or even years after removal of fuel from the nuclear reactor is of interest and only the long-lived fission products need be considered.

It has been found that the gamma energy released from fission products is relatively independent of the energy of the neutrons causing the fissions. However, the gamma-ray energy released and the photon energy spectrum depend significantly on the fissioning isotope, particularly in the first 10 s after fission. Generally, fissioning isotopes having a greater proportion of neutrons to protons produce fission-product chains of longer average length, with isotopes richer in neutrons and hence with greater available decay energy. Also, the photon energy spectrum generally becomes less energetic as the time after fission increases.

For very approximate calculations, the energy spectrum of delayed gamma photons from the fission of ^{235}U, at times up to about 500 s, may be approximated by the proportionality $N(E) \sim e^{-1.1E}$, where $N(E)$ is the delayed gamma yield (photons MeV^{-1} fission^{-1}) and E is the photon energy in MeV. The time dependence for the total gamma photon energy emission rate $F(t)$ (MeV s^{-1} fission^{-1}) is often described by the simple decay formula $F(t) = 1.4\ t^{-1.2}$, $10\ \mathrm{s} < t < 10^7$ s, where t is in seconds. More complicated (and accurate) expressions for $F(t)$ have been obtained from fits to experimental data; but for preliminary calculations the simpler result is usually adequate. It is observed that both ^{235}U and ^{239}Pu have roughly the same total gamma-ray-energy decay characteristics for up to 200 days after fission, at which time ^{235}U products begin to decay more rapidly until at 1 year after fission, the ^{239}Pu gamma activity is about 60% greater than that of ^{235}U.

For accurate calculations involving fission products, the variation with time after fission of the energy spectra of the photons must be taken into account. Often the energy spectra are averaged over discrete energy intervals and the energy emission rate in each energy group is considered as a function of time after fission. Computer codes, based on extensive libraries of radionuclide data, have been developed to compute the abundances and decay rates of the hundreds of fission-product radionuclides. An example of such calculations is shown in Fig. 3.



Radiation Sources. Figure 3
Total gamma-ray (G) and beta-particle (B) energy emission rates as a function of time after the thermal fission of ^{235}U. The curves identified by the numbers 1–6 are gamma emission rates for photons in the energy ranges 5–7.5, 4–5, 3–4, 2–3, 1–2, and 0–1 MeV, respectively

Capture Gamma Photons

The compound nucleus formed by neutron absorption is initially created in a highly excited state with excitation energy equal to the kinetic energy of the incident neutron plus the neutron binding energy, which averages about 7 MeV. The decay of this nucleus, usually within 10^{-12} s, and usually by way of intermediate states, typically produces several energetic photons. Generally, the probability a neutron causes an (n,γ) reaction is greatest for slow-moving *thermal neutrons*, that is, neutrons whose speed is in equilibrium with the thermal motion of the atoms in a medium. At high energies, it is more likely that a neutron scatters, thereby losing some of its kinetic energy, and then slows toward thermal energies.

Capture photons may be created intentionally by placing a material with a high thermal-neutron (n,γ) cross section in a thermal neutron beam. The energy spectrum of the resulting capture gamma photons can then be used to identify trace elements in the sample. More often, however, capture gamma photons are an undesired secondary source of radiation.

Inelastic Scattering Photons

The excited nucleus formed when a neutron is inelastically scattered decays to the ground state within about 10^{-14} s, with the excitation energy being released via one or more photons. Because of the constraints imposed by the conservation of energy and momentum in all scattering interactions, inelastic neutron scattering cannot occur unless the incident neutron energy is greater than $(A + 1)/A$ times the energy required to excite the scattering nucleus to its first excited state. Except for the heavy nuclides, neutron energies above about 0.5 MeV are typically required for inelastic scattering.

The detailed calculation of secondary photon source strengths from inelastic neutron scattering requires knowledge of the fast-neutron fluence, the inelastic scattering cross sections, and spectra of resultant photons, all as functions of the incident neutron energy. The cross sections and energy spectra of the secondary photons depend strongly on the incident neutron energy and the particular nuclide. Such inelastic scattering data are known only for the more important structural and shielding materials, and even the known data require extensive data libraries. Fortunately, in most situations, these secondary photons are of little importance compared to the capture photons. Although inelastic neutron scattering is usually neglected with regard to its secondary-photon radiation, it is a very important mechanism in the attenuation of fast neutrons, better even than elastic scattering in some cases.

Activation Photons

For many materials, absorption of a neutron produces a radionuclide with a half-life ranging from a fraction of a second to many years. The radiation produced by the subsequent decay of these activation nuclei may be very significant for materials that have been exposed to large neutron fluences, especially structural components in a reactor or accelerator. Many radionuclides encountered in research laboratories, medical facilities, and industry are produced as activation nuclides from neutron absorption in some parent material (see Table 4). Such nuclides decay, usually by beta emission, leaving the daughter nucleus in an excited state, which usually decays quickly to its ground state with the emission of one or more gamma photons. Thus, the apparent half-life of the photon emitter is that of the parent (or activation nuclide), while the number and energy of the photons are characteristic of the nuclear structure of the decay daughter.

Although most activation products of concern in shielding problems arise from neutron absorption, there is one important exception in water-moderated nuclear reactors. The ^{16}O in the water can be transmuted to ^{16} N in the presence of fission neutrons by an (n,p) reaction with a threshold energy of 9.6 MeV. ^{16}N decays with a 7.4-s half-life emitting gamma photons of 6.13 and 7.10 MeV (yields of 0.69 and 0.05 per decay). This gamma-ray source is very important in coolant channels of power reactors.

Positron Annihilation Photons

Positrons, generated either from the positron decay of radionuclides or from pair production interactions induced by high-energy photons, slow down in matter within about 10^{-10} s and are subsequently annihilated with electrons. With rare exception, the rest-mass

Radiation Sources. Table 4 Important radioisotopes produced by reactors and accelerators for use in medical, research, and industrial applications. Those isotopes commercially available for medical use are shown in bold. Decay data from NUDAT 2.5, National Nuclear Data Center, Brookhaven National Laboratory

Nuclide	Half-life	Decay modes[a]	Nuclide	Half-life	Decay modes[a]
^{3}H	12.33 years	β^- *	^{81m}Kr	13.1 s	EC IT
^{11}C	20.39 min	β^+ EC	^{85}Kr	10.76 years	β^-
^{13}N	9.965 min	β^+ EC	**^{82}Sr**[b]	25.6 days	EC *
^{14}C	5,730 years	β^- *	**^{89}Sr**	50.5 days	β^-
^{15}O	122.2 s	β^+ EC	**^{90}Sr**[c]	28.90 years	β^- *
^{18}F	109.8 min	β^+ EC	**^{99}Mo**[d]	65.94 h	β^-
^{22}Na	2.602 years	β^+ EC	^{103}Pd	16.99 days	EC
^{26}Al	7.17E5 years	β^+ EC	^{110m}Ag	249.8 days	β^- IT
^{28}Mg	20.91 h	β^-	**^{111}In**	2.80 days	EC
^{32}Si	153 years	β^- *	^{113m}In	99.48 min	IT
^{32}P	14.26 days	β^- *	**^{123}I**	13.27 h	EC
^{33}P	25.3 days	β^- *	**^{125}I**	59.40 days	EC
^{35}S	87.51 days	β^- *	**^{131}I**	8.025 days	β^-
^{36}Cl	3.01E5 years	$\beta^\pm$ EC	**^{133}Xe**	5.243 days	β^-
^{46}Sc	83.79 days	β^-	^{137}Cs[e]	30.1 years	β^- *
^{51}Cr	27.70 days	EC	^{140}La	1.679 days	β^-
^{54}Mn	312.1 days	β^- EC	^{148}Gd	70.9 years	α *
^{57}Co	271.7 days	EC	**^{153}Sm**	46.3 h	β^-
^{57}Cu	196 ms	β^+ EC	^{159}Gd	18.48 h	β^-
^{57}Cr	21.1 s	β^- *	^{169}Yb	32.02 days	EC
^{59}Fe	44.50 days	β^-	^{170}Tm	128.6 days	β^- EC
^{60}Co	5.271 years	β^-	^{186}Re	89.25 h	β^- EC
^{64}Cu	12.70 h	$\beta^\pm$ EC	^{191}Os	15.4 days	β^-
^{65}Zn	244.1 days	β^+ EC	^{192}Ir	73.83 days	β^- EC
^{67}Ga	3.261 days	EC	^{198}Au	2.696 days	β^-
^{68}Ga	67.71 min	β^+ EC	**^{201}Tl**	73 h	EC
^{75}Se	119.8 days	EC	^{204}Tl	3.78 years	β^- EC *
^{81}Rb	4.570 h	β^+ EC	^{210}Pb	22.2 years	$\alpha\beta^-$
^{82}Rb	1.273 min	β^+ EC	^{241}Am	432.6 years	α

[a] Decays without any gamma photon emission are denoted by *

[b] In equilibrium with decay product ^{32}Rb (1.273 min, β^+ EC)

[c] In equilibrium with decay product ^{90}Y (64 h, β^-)

[d] In equilibrium with decay product ^{99m}Tc (6.01 h, IT)

[e] In equilibrium with decay product ^{137m}Ba (2.552 min, IT)

Source: References [4, 5].

energy of the electron and positron is emitted in the form of two annihilation photons, each of energy $m_e c^2$ (= 0.511 MeV).

Sources of X-rays

The interaction of photons or charged particles with matter leads inevitably to the production of secondary x-ray photons. The x-rays in many applications have energies $\lesssim$100 keV, and hence are easily attenuated by any shield adequate for the primary radiation. Consequently, the secondary x-rays are often completely neglected in analyses involving higher-energy photons. There are many cases, though, when the energies of x-rays and Auger electrons must be accounted for as well as those of the x-rays. An example is the evaluation of radiation dose to the total body, or an organ of the body, after a radionuclide intake. This is a situation in which the source and receiver volumes may be the same.

There are important situations in which x-ray production is the only source of photons. To estimate the intensity, energies, and doses from the x-ray photons, it is necessary to understand how the x-rays are produced and some characteristics of the production mechanisms. There are two principal methods whereby secondary x-ray photons are generated: the rearrangement of atomic electron configurations leads to characteristic x-rays, and the deflection of charged particles in the nuclear electric field results in bremsstrahlung.

Characteristic X-rays and Fluorescence

The electrons around a nucleus are arranged in shells or layers, each of which can hold a maximum number of electrons. The two electrons in the innermost shell (*K* shell) are the most tightly bound, the six electrons in the next shell (*L* shell) are the next most tightly bound, and so on outward for the *M*, *N*,. . . shells. If the normal electron arrangement around a nucleus is altered, say by ejection of an inner electron, the electrons begin a complex series of transitions to vacancies in the inner shells (thereby acquiring higher binding energies) until the unexcited state of the atom is achieved. In each electronic transition, the difference in binding energy between the final and initial states is either emitted as a photon, called a *characteristic x-ray*, or given up to another electron which is ejected from the atom, called an *Auger electron.* The discrete electron energy levels and the transition probabilities between levels vary with the *Z* number of the atom, and thus the characteristic x-rays provide a unique signature for each element.

The number of x-rays with different energies is greatly increased by the multiplicity of electron energy levels available in each shell (1, 3, 5, 7, . . . distinct energy levels for the *K*, *L*, *M*, *N*, . . . shells, respectively). To identify the various characteristic x-rays for an element, many different schemes have been proposed. One of the more popular uses the letter of the shell whose vacancy is filled together with a numbered Greek subscript to identify a particular electron transition (e.g., $K_{\alpha 1}$ and $L_{\gamma 5}$).

Production of Characteristic X-Rays There are several methods whereby atoms may be excited and characteristic x-rays produced. A photoelectric absorption leaves the absorbing atom in an ionized state. If the incident photon energy is sufficiently greater than the binding energy of the *K*-shell electron, which ranges from 14 eV for hydrogen to 115 keV for uranium, it is most likely (80–100%) that a vacancy is created in the *K* shell and thus that the *K* series of x-rays dominates the subsequent secondary radiation. These x-ray photons produced from photoelectric absorption are often called *fluorescent radiation* and are widely used to identify trace elements in a sample by bombarding the sample with low-energy photons from a radioactive source or with x-rays from an x-ray machine and then observing the induced fluorescent radiation.

Characteristic x-rays can also arise following the decay of a radionuclide. In the decay process known as *electron capture*, an orbital electron, most likely from the *K* shell, is absorbed into the nucleus, thereby decreasing the nuclear charge by one unit. The resulting *K*-shell vacancy then gives rise to the *K* series of characteristic x-rays. A second source of characteristic x-rays which occurs in many radionuclides is a result of *internal conversion.* Most daughter nuclei formed as a result of any type of nuclear decay are left in excited states. This excitation energy may be either emitted as a gamma photon or transferred to an orbital electron which is ejected from the atom. Again, it is most likely that a *K*-shell electron is involved in this internal conversion process.

X-Ray Energies To generate a particular series of characteristic x-rays, an electron vacancy must be created in an appropriate electron shell. Such vacancies are created only when sufficient energy is transferred to an electron in that shell so as to allow it to break free of the atom or at least be transferred to an energy level above all the other electrons. The characteristic x-rays emitted when electrons fill a vacancy in a shell always have less energy than that required to create the vacancy. The most energetic x-rays arise from an electron filling a K-shell vacancy and, since the binding energy of K-shell electrons increases with the atomic number Z, the most energetic x-rays are K-shell x-rays from heavy atoms. For example, the K_α x-ray energy varies from only 0.52 keV for oxygen ($Z = 8$) to 6.4 keV for iron ($Z = 26$) to 98 keV for uranium ($Z = 92$). By comparison, the L series of x-rays for uranium occurs at energies around 15 keV. Thus, in most shielding situations, only the K series of x-rays from heavy elements are sufficiently penetrating to be of concern.

X-Ray Yields The *fluorescent yield* of a material is the fraction of the atoms with a vacancy in an inner electron shell that emit an x-ray upon the filling of the vacancy. The fluorescent yield increases dramatically with the Z number of the atom. For example, the fluorescent yield for vacancies in the K shell increases from 0.0069 for oxygen ($Z = 8$) to 0.97 for uranium ($Z = 92$). Thus, the secondary fluorescent radiation is of more concern for heavy materials.

Bremsstrahlung

A charged particle gives up its kinetic energy either by collisions with electrons along its path or by photon emission as it is deflected, and hence accelerated, by the electric fields of nuclei. The photons produced by the deflection of the charged particle are called *bremsstrahlung* (literally, “braking radiation”). For a given type of charged particle, the ratio of the rate at which the particle loses energy by bremsstrahlung to that by ionizing and exciting the surrounding medium is

$$\frac{\text{Radiation loss}}{\text{Ionization loss}} \simeq \frac{EZ}{700}\left(\frac{m_e}{M}\right)^2,$$

where E is in MeV, m_e is the electron mass, and M is the mass of the charged particle. From this result, it is seen that bremsstrahlung is more important for high-energy particles of small mass incident on high-Z material. In shielding situations, only electrons ($m_e/M = 1$) are ever of importance for their associated bremsstrahlung. All other charged particles are far too massive to produce significant amounts of bremsstrahlung. Bremsstrahlung from electrons, however, is of particular radiological interest for devices that accelerate electrons, such as betatrons and x-ray tubes, or for situations involving radionuclides that emit only beta particles.

Energy Distribution of Bremsstrahlung The energy distribution of the photons produced by the bremsstrahlung mechanism is continuous up to a maximum energy corresponding to the maximum kinetic energy of the incident charged particles. The exact shape of the continuous bremsstrahlung spectrum depends on many factors, including the energy distribution of the incident charged particles, the thickness of the target, and the amount of bremsstrahlung absorbed in the target and other masking material.

For monoenergetic electrons of energy E_o incident on a target thick compared to the electron range, the number of bremsstrahlung photons of energy E, per unit energy and per incident electron, emitted as the electron is completely slowed down can be approximated by the Kramer distribution

$$N(E_\text{o}, E) \simeq 2kZ\left(\frac{E_\text{o}}{E} - 1\right), \qquad E \leq E_\text{o},$$

where $k \simeq 0.0007\ \text{MeV}^{-1}$ is a normalization constant. The fraction of the incident electron’s kinetic energy that is subsequently emitted as bremsstrahlung can then be calculated from this approximation as kZE_o, which is usually a small fraction. For example, about 10% of the energy of a 2-MeV electron, when stopped in lead, is converted into bremsstrahlung.

Angular Distribution of Bremsstrahlung The angular distribution of bremsstrahlung is generally quite anisotropic and varies with the incident electron energy. Bremsstrahlung induced by low-energy electrons ($\lesssim$100 keV) is emitted over a relatively broad range of directions around the direction of the incident electron. As the electron energy increases, the direction of the peak intensity shifts increasingly toward the forward direction until, for electrons above a few

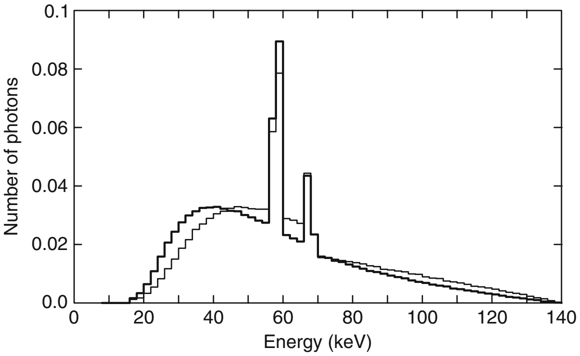
MeV, the bremsstrahlung is confined to a very narrow forward beam. The angular distribution of radiation leaving a target is very difficult to compute since it depends on the target size and orientation. For thin targets, the anisotropy of the bremsstrahlung resembles that for a single electron–nucleus interaction, while for thick targets multiple electron interactions and photon absorption in the target must be considered.

X-ray Machines The production of x-ray photons as bremsstrahlung and fluorescence occurs in any device that produces high-energy electrons. Devices that can produce significant quantities of x-rays are those in which a high voltage is used to accelerate electrons, which then strike an appropriate target material. Such is the basic principle of all x-ray tubes used in medical diagnosis and therapy, industrial applications, and research laboratories.

Although there are many different designs of x-ray sources for different applications, most designs for low-to-medium voltage sources ($\lesssim$180 kV) place the electron source (cathode) and electron target (anode) in a sealed glass tube. The glass tube acts as both an insulator between the anode and cathode and a chamber for the necessary vacuum through which the electrons are accelerated. The anodes of x-ray tubes incorporate a suitable metal upon which the electrons impinge and generate the bremsstrahlung and characteristic x-rays. Most of the electron energy is deposited in the anode as heat rather than being radiated away as x-rays, and thus heat removal is an important aspect in the design of x-ray tubes. Tungsten is the most commonly used target material because of its high atomic number and because of its high melting point, high thermal conductivity, and low vapor pressure. Occasionally, other target materials are used when different characteristic x-ray energies are desired. For most medical and dental diagnostic units, voltages between 40 and 150 kV are used, while medical therapy units may use 6–150 kV for superficial treatment or 180 kV to 50 MV for treatment requiring very penetrating radiation.

The energy spectrum of x-ray photons emitted from an x-ray tube has a continuous bremsstrahlung component up to the maximum electron energy, that is, the maximum voltage applied to the tube. If the applied voltage is sufficiently high as to cause ionization in the target material, there will also be characteristic x-ray lines superimposed on the continuous bremsstrahlung spectrum. Absorbing filters are used to minimize low-energy x-rays, which are damaging to skin. As the beam filtration increases, the low-energy x-rays are preferentially attenuated and the x-ray spectrum *hardens* and becomes more penetrating. These phenomena are illustrated in Fig. 4. Calculated exposure spectra of x-rays are shown for the same operating voltage but for two different amounts of beam filtration. As the filtration increases, lower energy x-rays are preferentially attenuated; the spectrum hardens and becomes more penetrating. Readily apparent in these spectra are the tungsten $K_{\alpha1}$ and $K_{\alpha2}$ characteristic x-rays.

The characteristic x-rays may contribute a substantial fraction of the total x-ray emission. For example, the *L*-shell radiation from a tungsten target is between 20% and 35% of the total energy emission when voltages between 15 and 50 kV are used. Above and below this voltage range, the *L* component rapidly decreases in importance. However, even a small degree of filtering of the x-ray beam effectively eliminates the low-energy portion of the spectrum containing the *L*-shell x-rays. The higher-energy *K*-series x-rays from a tungsten target contribute a maximum of 12% of the



Radiation Sources. Figure 4
Measured photon spectra from a Machlett Aeromax x-ray tube (tungsten anode) operated at a constant 140 kV potential. This tube has an inherent filter thickness of 2.50-mm aluminum equivalent and produces the spectrum shown by the thick line. The addition of an external 6-mm aluminum filter hardens the spectrum shown by the thin line. Both spectra are normalized to unit area. Data are from [6]

x-ray emission from the target for operating voltages between 100 and 200 kV.

Synchrotron Photons

When a charged particle moving in a straight line is accelerated by deflecting it in an electromagnetic field, the perturbation in the particle's electric field travels away from the particle at the speed of light and is observed as electromagnetic radiation (photons). Such is the origin of bremsstrahlung produced when fast electrons (beta particles) are deflected by the electric field of a nucleus.

This same mechanism can be used to produce intense photon radiation by deflecting an electron beam by magnetic fields. In a special accelerator called a *synchrotron*, highly relativistic electrons are forced to move in a circular path inside a storage ring by placing *bending magnets* along the ring. Photons are emitted when the beam is accelerated transversely by (1) the bending magnets (used to form the circular electron beam), and by (2) insertion device magnets such as *undulators*, *wigglers*, and *wavelength shifters*.

Because the electrons are very relativistic, the synchrotron radiation is emitted in a very narrow cone in the direction of electron travel as they are deflected. Undulators cause the beam to be deflected sinusoidally by a weak oscillatory magnetic field, thereby producing nearly monochromatic photons. By contrast, a wiggler uses a strong oscillatory magnetic field which, because of relativistic effects, produces distorted sinusoidal deflections of the electron beam and synchrotron radiation with multiple harmonics, that is, a line spectrum. If very strong magnetic fields are used, many harmonics are produced that merge to yield a continuous spectrum ranging from the infrared to hard x-rays. By placing undulators or wigglers at a specific location in the storage ring, very intense and narrowly collimated beams of photons with energies up to a few keV can be produced to use, for example, in x-ray diffraction analysis.

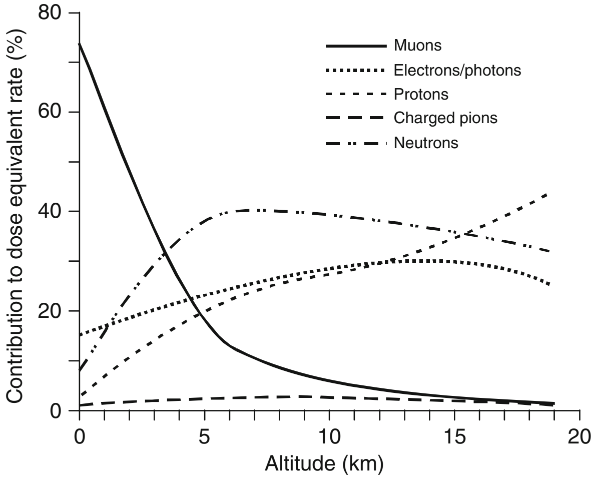
Cosmic Rays, Solar Radiation, and Trapped Radiation Belts

The earth is subjected continuously to radiation with sources in our sun and its corona, from sources within our galaxy, and from sources beyond our galaxy. In addition, surrounding the earth are belts of trapped particles with solar origins. Radiation reaching the earth's atmosphere consists of high-energy electrons and atomic nuclei. Hydrogen nuclei (protons) constitute the major component, with heavier atoms decreasing in importance with increasing atomic number. The highest energy particles originate in our galaxy and more distant galaxies and are referred to as galactic cosmic radiation (GCR). Cascades of nuclear interactions in the atmosphere give rise to many types of secondary particles. Of much lower energy are particles of solar origin, which are highly variable in time and are associated with solar activity. Sources of these particles are sometimes associated with solar flares but are more generally identified with solar particle events (SPE) or coronal mass ejections (CME). The number of solar flare events and SPE emissions fluctuate with the 11-year cycle associated with solar activity. GCR intensity is modulated by SPE emissions, being minimal when solar activity is maximal.

Galactic Cosmic Radiation

At the earth's surface, cosmic radiation dose rates are largely due to muons and electrons. The intensity and angular distribution of galactic radiation reaching the earth is affected by the earth's magnetic field and perturbed by magnetic disturbances generated by solar flare activity. Consequently, at any given location, cosmic ray doses may vary in time by a factor of 3. At any given time, cosmic ray dose rates at sea level may vary with geomagnetic latitude by as much as a factor of 8, being greatest at the pole and least at the equator. Cosmic ray dose rates also increase with altitude. At geomagnetic latitude 55°N, for example, the absorbed dose rate in tissue approximately doubles with each 2.75 km (9,000 ft) increase in altitude, up to 10 km (33,000 ft). Figure 5 illustrates the relative importance, in terms of dose rate, for cosmic rays and their reaction products in the atmosphere. Cosmic ray dose rates affecting populations vary strongly with latitude. Table 5 describes this variation. The outdoor and indoor average effective dose rates for space radiation in the most heavily populated urban areas in the USA are 45 and 36 nSv/h [7]. Population averaged annual doses are about the same in the northern and southern hemispheres, and globally amount to 31 nSv/h for charged particles plus 5.5 nSv/h for neutrons. GCR energies span a vast range and there is no way of shielding astronauts from GCR effects. On the earth's surface, the GCR presence results in a source of steady low-dose-rate radiation.

As a result of nuclear reactions of cosmic rays with constituents of the atmosphere, secondary neutrons, protons, and pions, mainly, are produced. Subsequent pion decay results in electrons, photons, and muons. Muon decay, in turn, leads to secondary electrons, as do scattering interactions of charged particles in the atmosphere. Cosmic ray debris that reaches the surface of the earth consists mainly of muons and electrons with a few neutrons. Except for short-term influences of solar activity, galactic cosmic radiation has been constant in intensity for at least several thousand years. The influence of solar activity is cyclical and the principal variation is on an 11-year cycle. The geomagnetic field of the earth is responsible for limiting the number of cosmic rays that can reach the atmosphere thus accounting for the strong effect of latitude on cosmic-ray dose rates.



Radiation Sources. Figure 5
Components of the dose equivalent rate from cosmic rays in the atmosphere (Reproduced from [2], derived from [5])

Solar Particle Events

Both SPE and CME emissions are mainly hydrogen and helium nuclei, that is, protons and alpha particles, predominantly the former. Electrons are thought to be emitted as well, but with energies less than those of protons by a factor equal to the ratio of the rest masses. Energy spectra are highly variable, as are temporal variations of intensity. A typical course of events for a flare is as follows. Gamma and x-ray emission takes place over about 4 h as is evidenced by radio interference. The first significant quantities of protons reach the earth after about 15 h and peak proton intensity occurs at about 40 h after the solar eruption.

Solar particle events are closely related to solar flares associated with sunspots with intense magnetic fields linking the corona to the solar interior. CME emissions are not directly connected to flares, but originate in the corona driven by the energy of the sun's

Radiation Sources. Table 5 Cosmic ray dose rate variation at sea level as a function of latitude for the northern and southern hemispheres

		Effective dose rate (nSv/h)	
Latitude (deg N–S)	Population% in latitude band (N–S)	Charged particles	Neutrons
60–70	0.4–0	32	10.9
50–60	13.7–0.5	32	10
40–50	15.5–0.9	32	7.8
30–40	20.4–13	32	5.3
20–30	32.7–14.9	30	4
10–20	11–16.7	30	3.7
0–10	6.3–54	30	3.6

Source: Reference [2].

magnetic field. While of too low energy to contribute to radiation doses at the surface of the earth, these radiations, which fluctuate cyclically with an 11-year period, perturb earth's magnetic field and thereby modulate galactic cosmic-ray intensities with the same period. Maxima in solar flare activity lead to minimal GCR intensity. SPE and CME emissions, in comparison to galactic cosmic rays, are of little significance as a hazard in aircraft flight or low orbital space travel. On the other hand, these radiations present considerable, life-threatening risk to personnel and equipment in space travel outside the earth's magnetic field. Protection of astronauts in space missions beyond low-earth orbit is addressed in [8].

Trapped Radiation Belts

Released continuously from the sun, as an extension of the corona, is the solar wind, a plasma of low-energy protons and electrons. The solar wind does not present a radiation hazard, even in interplanetary space travel. However, it does affect the interplanetary magnetic field and the shape of the geomagnetically trapped radiation belts. These radiation belts are thought to be supplied by captured solar-wind particles and by decay into protons and electrons of neutrons created by interactions of galactic cosmic rays in the atmosphere. The trapped radiation can present a significant hazard to personnel and equipment in space missions.

The earth's geomagnetically trapped radiation belts are also known as Van Allen belts in recognition of James A. Van Allen and his coworkers who discovered their existence in 1958. There are two belts. The inner belt consists of protons and electrons, the protons being responsible for radiation doses in the region. The outer belt consists primarily of electrons. The particles travel in helical trajectories determined by the magnetic field surrounding the planet. They occur at maximum altitude at the equator and approach the earth most closely near the poles. At the equator, the inner belt extends to about 2.8 earth radii. The center of the outer belt is at about 5 earth radii. The solar wind compresses the trapped radiation on the sunny side of the earth and the compression is enhanced by solar flare activity. In the earth's shadow, the belts are distended as the solar wind sweeps the magnetosphere outward. In a plane through the earth, perpendicular to the earth–sun axis, the proton and electron belts are maximum in intensity at altitudes of about 3,000 and 18,000 km, respectively.

In the southern Atlantic Ocean, there is an eccentricity of the geomagnetic field with respect to the earth's center, and magnetic field lines dip closer to earth. This region, the South Atlantic Anomaly, is the primary source of radiation exposure to astronaut crew members in low-altitude and low-inclination missions. Radiation protection guidance for low-earth orbit missions is found in [9] and [10].

Radiation Sources Used in Human Activities

Life on earth is continually subjected to radiation of natural origin. Exposure is from sources outside the body, arising from cosmic radiation and radionuclides in the environment, and from sources inside the body, arising from ingested or inhaled radionuclides retained in the body. Natural sources are the major contributors to human radiation exposure and represent a reference against which exposure to man-made sources may be compared. Table 6 summarizes radiation doses to man resulting from natural sources. Listed in the table are both doses to individual organs or tissues of the body and the effective dose equivalent, which is a composite dose weighted by the relative radiation sensitivities of many organs and tissues of the body.

Since the early 1980s, there has been negligible change in exposure to naturally occurring radiation, an increase estimated from 3 to 3.1 mSv annually. However, in the USA, medical diagnostic exposures have increased by a factor of 5.5 by 2006, an increase from 0.53 to 3 mSv annually. Of the 3 mSv total, 1.47 mSv is from computed tomography (CT) scans, 0.43 mSv for interventional fluoroscopy, 0.77 mSv for nuclear medicine procedures, and 0.33 mSv for conventional radiography and fluoroscopy (10).

Sources in Medicine

Very shortly after their discoveries at the end of the nineteenth century, radium and x-rays were used for medical purposes – radium sources being concentrated from natural materials and x-rays being generated using new technology. These were the only radiation sources seeing significant use until the 1930s, when

Radiation Sources. Table 6 Summary of US annual doses from natural background radiation

	Average annual dose equivalent (mrem)				
Radiation source	Bronchial epithelium	Other soft tissues	Bone surfaces	Bone marrow	Effective dose equivalent
Cosmic radiation	27	27	27	27	27
Cosmogenic nuclides	1	1	1	3	1
External terrestrial	28	28	28	28	28
Inhaled nuclides	2,400				200
Nuclides in the body	36	36	110	50	39
Totals (rounded)	2,500	90	170	110	300

Source: Reference [11].

research into nuclear fission began and when high-energy particle accelerators were developed for nuclear research. In the first half of the twentieth century, x-rays revolutionized diagnostic medicine. In the second half, accelerator radiation and radionuclides produced by accelerators and nuclear reactors established radiography, radiation therapy, and nuclear medicine, both diagnostic and therapeutic, as mature medical sciences. Table 4 lists the radioisotopes commonly used in medicine and industry. Some of these radionuclides are produced in nuclear reactors, either as products of fission or as products of neutron absorption. Nuclei of these isotopes are rich in neutrons and tend to decay by emission of negative beta particles, thereby becoming more positive in charge and more stable. Other isotopes are produced in accelerators. These generally have nuclei deficient in neutrons and tend to decay either by emission of a position or capture of an electron, either process leaving the nucleus more negative and more stable.

There are three broad categories of medical procedures resulting in human radiation exposure: (1) diagnostic x-ray examinations, including mammography and computed tomographic (CT) scans, (2) diagnostic nuclear medicine, and (3) radiation therapy.

Diagnostic X-Rays Of all the radiation exposures to the general public arising from human activity, the greatest is due to medical procedures, and collective exposures from diagnostic x-rays dominate all other medical exposures. Also, the population subgroup receiving diagnostic x-rays is not small. In the USA, about 250 million medical x rays are delivered annually, as are about 70 million CT scans. About 900 thousand persons receive radiation therapy annually [2, 7].

Diagnostic Nuclear Medicine Internally administered radionuclides are used medically for imaging studies of various body organs and for non-imaging studies such as thyroid uptake and blood volume measurements. Such uses present hazards for both patients and medical staff. Radiopharmaceuticals are also used for in vitro studies such as radioimmunoassay measurements and thus are of potential hazard to medical staff. Frequencies of procedures, while steadily increasing, vary widely from country to country. As of 2000, in industrialized countries, about 10–40 examinations involving radiopharmaceuticals are carried out annually per 1,000 population. In developing countries, annual frequencies are on the order of 0.2–2 examinations per 1,000 population. In the USA in 2006, for example, some 18 million radionuclide administrations were performed annually for diagnostic purposes [2, 7].

Radiation Therapy There are three broad categories of radiation therapy– teletherapy, brachytherapy, and therapy using administered radiation sources. *Teletherapy* involves external beams from sources such as sealed ^{60}Co sources, x-ray machines, and accelerators that generate electron, proton, neutron, or x-ray beams. *Brachytherapy* involves sources placed within body cavities (*intracavitary* means) or placed directly within tumor-bearing tissue (*interstitial* means). In the USA, Europe, and Japan, the frequencies for teletherapy and brachytherapy procedures exceed 2,000 annually per million population.

Thyroid disorders, including cancer, for many years have been treated by ^{131}I, usually by oral administration. Introduced about 1980, in association with the development of techniques for producing monoclonal antibodies, were new cancer diagnosis and treatment methodologies called radioimmunoimaging and radioimmunotherapy. The therapy involves administration of large doses of antibodies tagged with radionuclides and selected to bind with antigens on the surfaces of tumor cells. Imaging involves administration of very much smaller doses, with the goal of detecting the presence of tumor cells using standard camera and scanner imaging techniques. Imaging requires the use of radionuclides such as ^{99m}Tc, which emit low-energy gamma rays. Therapy involves the use of radionuclides emitting beta particles and electrons, with minimum emission of gamma rays, thus limiting radiation exposure, to the extent possible, to tumor cells alone. Among radionuclides used in radioimmunotherapy are ^{75}Se, ^{90}Y, ^{111}In, ^{125}I, ^{186}Re, and ^{191}Os.

Occupational Medical Exposure A world survey conducted by the United Nations for the years 1990–1994 reports the annual average effective dose 1.39 mSv to some 550,000 workers receiving measurable doses (2.3 million total). Of this group, the greatest number were involved in diagnostic radiology (350,000 at 1.34 mSv). Overall, exposures were in the range of 0.9–1.7 mSv annually [2].

Accelerator Sources

The earliest particle accelerators were the x-ray tubes of the late nineteenth century. Indeed, the radio and television (cathode-ray) tubes of the twentieth century are low-voltage electron accelerators. As electrons beams are stopped, x-rays are produced, inadvertently in the case of radio tubes, and deliberately in the case of x-ray generators.

Modern charged-particle accelerators date from the early 1930s, when Cockroft and Walton in England, and Lawrence and Livingston in America developed particle accelerators for research purposes using beams of electrons or ions. Over the years, steady advances have been made in types of accelerators, in the energies of the particles accelerated, and in the magnitude of the current carried by the charged particle beams. Accelerators continue to serve at the frontiers of atomic and nuclear physics as well as the materials sciences. Moreover, accelerators play an ever more important role in diagnostic and therapeutic medicine and in industrial production processes such as radiography, analysis of materials, radiation processing, and radioisotope production.

Particle accelerators may be classified technologically as direct (potential drop) accelerators and indirect (radio-frequency, plasma) accelerators. Among the former are the Van de Graaff and Cockroft-Walton devices. Among the latter are linear accelerators, betatrons, cyclotrons, and synchrotrons. In the linear accelerator, the particles travel in straight lines, accelerated by the electric fields along their paths. In cyclic accelerators, magnets are used to direct particles into approximately circular paths, along which they may pass through the same accelerating electric fields many times along their paths. The ultimate energies reached by the accelerated particles have increased from about 10^6 eV in the accelerators of the 1930s to 10^{12} eV in modern research accelerators.

By their very nature and function, accelerators are intense radiation sources. In certain applications, accelerated beams of charged particles are extracted from accelerators and directed onto external receivers. Medical applications and radiation processing see this use of accelerators. In other applications, charged particle beams impinge on internal target receivers designed to act as desired sources of secondary radiations such as x-rays or neutrons. In all cases, beams are stopped by targets within which secondary x-rays, neutrons, and other particles such as mesons may be produced as undesirable but unavoidable by-product radiations. Radiation shielding integral with the accelerator as well as structural radiation shielding surrounding the accelerator are necessary for personnel protection.

The production of secondary radiations arises mainly from three phenomena, direct nuclear reactions of ions or electron with accelerator components, electromagnetic cascades, and hadronic cascades.

Among the secondary radiations are neutrons, which in turn may be absorbed in accelerator and structural materials thereby leading to capture gamma rays and radioactive reaction products.

Representative of the direct nuclear reactions are those of relatively low-energy proton or deuteron beams in light-element targets. A popular method of generating energetic neutrons, for example, involves interactions of deuterons accelerated to 150 keV with tritium atoms in a target. The resulting reaction, $^{3}\mathrm{H}(d,n)^{4}\mathrm{He}$, produces an approximately isotropic and monoenergetic source of 14-MeV neutrons. Other such reactions include $^{3}\mathrm{H}(p,n)^{3}\mathrm{He}$, $^{2}\mathrm{H}(d,n)^{3}\mathrm{He}$, and $^{7}\mathrm{Li}(p,n)^{7}\mathrm{Be}$.

The electromagnetic cascade involves exchanges of the kinetic energy of an electron to electromagnetic energy of multiple photons in the bremsstrahlung process, followed by creation of the rest-mass and kinetic energies of an electron–positron pair in the pair-production process experienced by the photons. As the positrons and electrons lose kinetic energy radiatively, more photons are produced, and the cascade continues. The cascade is quenched when photons have insufficient energy to generate the rest masses of the electron–positron pair and when electron radiative energy losses fall below collisional energy losses.

In high-energy electron or proton accelerators, hadronic cascades may be produced when particles collide with atoms in the accelerator target or, inadvertently, with some other accelerator component, giving rise to many reaction products, including pions, kaons, protons, and neutrons. There is also exchange with the electromagnetic cascade via photodisintegration reactions and by production of energetic gamma rays upon decay of π^0 mesons. Propagation of the hadronic cascade occurs through reactions of the secondary protons and neutrons, and is especially important for nucleon energies of 150 MeV or greater. The cascade process produces most of the induced radioactivity at high-energy accelerators. Many reaction-product nuclei are in highly excited nuclear states and relax by emission of neutrons, whose subsequent absorption leads, in many cases, to radioactive by-products.

Water, plastics, and oils in the radiation environs of high-energy accelerators yield ^{7}Be and ^{11}C. Aluminum yields these same radionuclides plus ^{18}F, ^{22}Na, and ^{24}Na. Steel, stainless steel, and copper yield all the aforementioned, plus a very wide range of radionuclides, especially those of V, Cr, Mn, Co, Fe, Ni, Cu, and Zn. Neutron absorption in structural concrete also leads to a wide range of radionuclides, among which ^{24}Na is a major concern. This nuclide has a half-life of 15 h and, in each decay, emits high-energy beta particles and gamma rays.

Industrial Isotope Sources

Radionuclides used in industry contribute very little to collective population doses, although individual occupational exposures may be significant. The largest sources are those used in radiography, typically comprising 10–100 Ci of ^{192}Ir, ^{137}Cs, ^{170}Tm, or ^{60}Co. Borehole logging is accomplished using somewhat lower activity gamma-ray sources and neutron sources such as mixtures of plutonium, americium, or californium with beryllium. Much lower activity sources, often $^{90}\mathrm{Sr} - {}^{90}\mathrm{Y}$ beta-particle sources, are used for various instrumentation and gaging applications.

There are many consumer products containing radiation sources. While these sources are very weak and no one individual receives significant radiation exposure, many persons are involved. For example, the soil, water supplies, and building materials contain low concentrations of naturally occurring radionuclides. Electronic devices emit very low levels of x-rays, and devices ranging from luminous timepieces to smoke detectors contain weak radiation sources. Even the use of tobacco exposes smokers to alpha particles from naturally occurring ^{210}Po in the tobacco leaf.

Various modern technologies have led to human radiation exposures in excess of those which would have occurred in the absence of the technologies. For example, the mining of coal and other minerals and their use is responsible for increased releases of naturally occurring radionuclides to the environment. World production of coal is about 4 billion tonnes annually. About 70% is used in generation of electricity, the balance mainly in domestic heating and cooking. Coal contains ^{40}K, and the ^{238}U and ^{232}Th decay chains in widely varying concentrations. Depending on the nature of combustion, radionuclides are partitioned between fly ash and bottom ash. The smaller-sized fly ash particles are more heavily enriched with radionuclides, particularly ^{210}Pb and ^{210}Po. The average ash content of coal is about 10% by weight, but may be as high as 40%. Efficiency of ash removal in power plants is quite variable– from only 80% removal to as much as 99% removal, the average being about 97.5%. In domestic use of coal, as much as 50% of the total ash is released into the atmosphere. In terms of doses to individual tissues, the main impact of atmospheric releases during combustion is the dose to bone surface cells accruing from inhalation of ^{232}Th present in the downwind plumes of particulates from plants.

Annually, some 1.4 billion tonnes of phosphate rock are mined and processed for use in production of fertilizers and phosphoric acid. By-product (phospho)gypsum finds wide use in the construction industry. The USA produces about 38% of the phosphate rock, the former Soviet Union 19%, and Morocco 14%. Sedimentary phosphate rock contains high concentrations of radionuclides in the ^{238}U decay chain. Most airborne radioactivity releases are associated with dust produced in strip mining, grinding, and drying of the ore. Utilization of the phosphates leads to both internal and external radiation exposure, the greatest exposure resulting from use of by-product gypsum in construction.

The Nuclear Power Industry

In mid-2008, there were 439 nuclear power plants operating around the world, with a total electrical generating capacity of 372,000 MW. An additional 42 nuclear plants were under construction in 15 countries. Most of the generating capacity consisted of pressurized-water and boiling-water reactors, which use ordinary water as coolant. Gas-cooled reactors, heavy-water reactors, light-water graphite reactors, and sodium-cooled reactors provided the balance of the capacity.

Nuclear Power Reactors The fission process and production of neutrons associated with reactor operation lead to a wide array of radioactive fission products and activation products arising from neutron absorption. Moreover, large quantities of uranium and plutonium are fissioned in modern nuclear power plants (typically 3 kg/day) to produce the thermal energy needed to produce electricity. Consequently, large quantities of fission products are produced and accumulate in the fuel. Also contained within the fuel are actinides produced by cumulative neutron absorption in uranium, thorium, and plutonium fuels. The actinides are characterized by spontaneous fission in competition with alpha-particle decay, and require sequestration to the same degree as the fission products.

One way of categorizing the generated radionuclides is by their physical–chemical behavior, namely, (1) noble gases (2) ^{3}H and ^{14}C, (3) halogens, and (4) particulates. These divisions are based on the relative ease of isolation of the radionuclides from airborne effluents. The noble gases include the many isotopes of the krypton and xenon fission products as well as the activation product ^{41}Ar. These elements cannot be removed from a gas stream by filtration. Halogens include the many isotopes of the bromine and iodine fission products. If they are present in a gas stream, they are likely to be in a chemical form unsuitable for filtration, and effective removal requires adsorption on a material such as activated charcoal. Other radionuclides and the halogens in ionic form may be removed from a gas stream by filtration. In aqueous liquids, the particulates may be isolated by evaporation, filtration, or ion exchange. The halogens, unless in ionic form, cannot be isolated by evaporation or filtration, nor can noble gases. Special cases are ^{3}H in the form of tritiated water and ^{14}C as carbon dioxide. The tritium can be isolated only with very great difficulty, and CO_2 removal requires chemical treatment.

There are two sources of radionuclides in reactor coolant, leakage from defective fuel rods and activation products produced by neutron interactions in the coolant itself or with fuel and structure in contact with the coolant. Activation product sources are inevitable, and include a number of radionuclides which may be produced in the coolant. For example, ^{16}N is produced as a result of neutron interactions with oxygen, ^{41}Ar as a result of neutron absorption in naturally occurring argon in the atmosphere and ^{3}H as a result of neutron absorption in deuterium and, especially in pressurized-water reactors, by neutron-induced breakup of ^{10}B. Of course, in a sodium-cooled fast reactor, activation of natural sodium to short-lived ^{24}Na is an important consideration for in-plant radiation protection. Other activation products include isotopes of iron, cobalt, chromium, manganese, and other constituents of structural and special-purpose alloys. The radionuclides are leached into the coolant stream. They then may be adsorbed on surfaces or trapped as particulates in the boundaries of coolant streams within the plant, only later to be resuspended in the coolant. These sources can be minimized by carefully specifying the alloy and trace-element concentrations in plant components.

Uranium Mining and Fuel Fabrication In the preparation of new fuel for nuclear reactors, the radiation sources encountered are natural sources associated with the uranium and thorium decay chains.

The principal release of radiation sources associated with uranium mining, underground or open pit, is release of natural ^{222}Rn to the atmosphere. Airborne particulates containing natural uranium daughter products also arise from open pit mining and from ore crushing and grinding in the milling process. Mill tailings can also become a long-term source of radiative contamination due to wind and water erosion, leaching, and radon release, the degree depending on the tailing-stabilization program followed. Mining and milling operations are generally conducted in remote areas, and liquid releases containing dissolved uranium daughter products are of little impact on human populations.

The product of milling is U_3O_8 "yellow cake" ore concentrate. In this phase of the nuclear fuel cycle, the concentrate is purified and most often converted to UF_6 for enrichment in ^{235}U via gaseous diffusion or centrifuge processes. Prior to fuel fabrication, the uranium is converted to the metallic or the ceramic UO_2 form suitable for use in fuel elements. Large quantities of uranium depleted in ^{235}U are by-products of the enrichment process. Under current practice, the depleted uranium is held in storage as being potentially valuable for use in breeder reactors. In this stage of the nuclear fuel cycle, there are relatively minor liquid and gaseous releases of uranium and daughter products to the environment.

Fuel Reprocessing As nuclear fuel reaches the end of its useful life in power generation, there remain within the fuel recoverable quantities of uranium and plutonium which may be extracted for reuse in the fuel reprocessing stage of the nuclear fuel cycle. Whether or not the fuel is reprocessed is governed by economic and political considerations. Among the former are costs of reprocessing as compared to costs of mining, milling, conversion, and enrichment of new stocks of uranium. Among the latter are concerns over the potential diversion of plutonium to nuclear-weapons use.

In the reprocessing of oxide fuels, the spent fuel is first dissolved in nitric acid. Plutonium and uranium are extracted into a separate organic phase from which they are ultimately recovered and converted into the oxide form. The aqueous phase containing fission and activation products is then neutralized and stored in liquid form pending solidification and ultimate disposal. Because one reprocessing plant may serve scores of power plants, inventories of radionuclides in process may be very great and extraordinary design features and safety procedures are called for. Because of the time delays between removal of fuel from service and reprocessing, concerns are with only relatively long-lived radionuclides, notably ^{3}H, ^{14}C, ^{85}Kr, ^{90}Sr, ^{106}Ru, ^{129}I, ^{134}Cs, and ^{137}Cs.

During the dissolution step of reprocessing, all the ^{85}Kr, the bulk of the ^{14}C (as CO_2), and portions of the ^{3}H and ^{129}I appear in a gas phase. This gas is cleaned, dried, and released through a tall stack to the atmosphere. All the ^{85}Kr is thus released; however, the major part of the ^{3}H is removed in the drying process and the bulk of the ^{129}I and ^{14}C is removed by reaction with caustic soda. The ^{14}C may then be precipitated and held as solid waste. Depending on the degree of liquid-effluent cleanup, some of the ^{129}I and other fission products subsequently may be released to the environment.

Waste Storage and Disposal Wastes generated in the nuclear fuel cycle fall into the broad categories of high-level wastes (HLW) and low-level wastes (LLW). The former, comprising unprocessed spent fuel or liquid residues from fuel reprocessing, accounts for only about 1–5% of the waste volume, but about 99% of the waste activity. The latter is comprised of in-reactor components, filter media, ion-exchange resins, contaminated clothing and tools, and laboratory wastes. For the most part, LLW consists of short-lived beta-particle and gamma-ray emitters. Wastes of low specific activity, but containing long-lived alpha-particle emitters, for example, ^{239}Pu, require special handling more in the nature of that required for HLW.

In the USA, fuel elements from commercial reactors are presently not processed. By the year 2000, the cumulative spent fuel reached about 16,000 m^3, amounting to 40,000 t of uranium and fission products. Most of this spent fuel will be stored at the plant sites where it is generated which are primarily in eastern states.

Nuclear Power and Occupational Exposure Table 7 summarizes a United Nations survey of occupational exposures as well as public exposures based on nuclear power operations in the 1990s. Improvements continue to be made and US occupational annual committed occupational doses have by 2007 declined from about 3 to 1.1 person Sv per GWy(e).

Nuclear Explosives

Large fractions of radioactive debris from atmospheric nuclear weapons tests are distributed globally, and the radionuclides remain in the biosphere indefinitely. The hazard is better characterized by the long-term dose commitment than by the dose rate at any instant and location. The fusion and fission energy released in a nuclear-weapon explosion is usually measured in units of megatons (Mt). One megaton refers to the release of 10^{15} cal of explosive energy– approximately the amount of energy released in the detonation of 10^6 metric tons of TNT. The quantity of fission products produced in a nuclear explosion is proportional to the *weapon fission yield*. For a 1-Mt weapon fission yield, there must be the complete fissioning of about 56 kg of uranium or plutonium. The quantities of ^{3}H and ^{14}C, which are produced in the atmosphere by interactions of high-energy fission neutrons, are also proportional to the *weapon fusion yield*. There are several fusion reactions used in thermonuclear devices, with a 1-Mt weapon fusion yield requiring, for example, the fusion of 7.4 kg of tritium with 4.9 kg of deuterium.

The disposition of weapon debris may be divided into three categories, local fallout, tropospheric fallout, and stratospheric fallout. Local fallout, comprising as much as 50% of the debris and consisting of large particles, is defined as that deposited within 100 miles of the detonation site. Depending on detonation altitude and weather conditions, a portion of the weapon's

Radiation Sources. Table 7 Collective doses incurred by the public and workers from the nuclear fuel cycle, normalized to unit electrical energy generation

	Committed dose (person Sv) per GWy(e)		
	General population		
Operation	Local/regional	Waste/global[a]	Occupational
Mining	0.19	–	1.72
Milling	0.008	–	0.11
Mine tailings	0.04	7.5	–
Fuel fabrication	0.003	–	0.1
Reactor operation	0.44	0.5	3.9
Fuel reprocessing	0.13	0.05	3.0
Transportation	<0.1	–	–
Global dispersion	–	40	–
Research	–	–	1.0
Enrichment	–	–	0.02
Total	0.91	50	9.8

[a] For solid waste and global dispersion, committed dose is for 10,000 years

Source: Reference [2]: (1995–1997 general, 1990–1994 occupational).

debris is injected into the stratosphere and a portion remains in the troposphere. These two atmospheric regions are separated by the tropopause (about 16 km altitude at the equator and 9 km at the poles). Temperature decreases with elevation in the troposphere. This hydrodynamically unstable condition leads to the development of convective weather patterns superimposed upon generally westerly winds. In the stratosphere, temperature is more nearly constant or, in equatorial regions, even rises with elevation. Vertical convective motion is relatively slight and the tropical temperature inversion restricts transfer of material in the stratosphere from hemisphere to hemisphere.

Debris in the troposphere is distributed in longitude but remains within a band of about 30° of latitude. The mean lifetime of radioactive debris in the troposphere is about 30 days and tropospheric fallout is important for radionuclides with half-lives of a few days to several months. Over the years, the bulk of the radioactive debris from weapons tests has been injected into the stratosphere in the northern hemisphere and at altitudes less than 20 km. Mechanisms for transfer of the debris to the troposphere and thence to fallout on the earth's surface are complex. At elevations less than 20 km, the half-life for transfer of aerosols between hemispheres is about 60 months, while the half-life for transfer to the troposphere is only about 10 months, with little material crossing the tropopause in equatorial regions. Consequently, the bulk of the fallout from any one test occurs over the hemisphere of injection and in temperate regions. In terms of the megatons of fission energy, in the period prior to 1980, 78% of the debris was injected into the stratosphere– 70% into the northern hemisphere and 8% into the southern.

Eight radionuclides contribute significantly to the committed effective dose equivalent to the population. These are ^{137}Cs, ^{131}I, ^{14}C, ^{239}Pu, ^{90}Sr, ^{106}Ru, ^{144}Ce, and ^{3}H. Because of its long half-life, 5,730 years, the commitment from ^{14}C extends over many human generations. The collective effective dose equivalent commitment into the indefinite future due to weapons tests to date is equivalent to about four extra years of exposure of the current world population to natural background radiation.

High-level radioactive wastes generated in the USA in the production of nuclear weapons have accumulated for decades. The wastes are stored at three sites, one in the state of Washington, one in Idaho, and one in South Carolina. The approximately 9,000 t of waste has a volume of 380,000 cubic meters and there are plans to dispose of this waste in a repository used also for disposal of spent fuel for nuclear power plants.

Future Directions

Accelerators

Research, materials processing, and teletherapy accelerators employ higher and higher energy beams and beam currents. The production of secondary radiations associated with beam interactions with targets and structure lead to radiation sources of new types and increased magnitudes. Proton-beam accelerators grow in use for specialized radiation therapy. Synchrotrons deliver beams of protons with energies up to 250 MeV, with increased demands for new and better radiation surveillance, shielding, and dosimetry. Minimization of radiation damage to accelerator components calls for more precise beam simulation in the design process as well as more robust components.

Space Activities

In low-earth orbit, galactic cosmic rays (GCR) properties are well known except for short-term enhancements caused by solar particle events (SPE), especially in polar regions [10]. Consequences may be important in extravehicular activities (EVA) outside the Space Shuttle or International Space Station. There is a need for real-time measurement of instantaneous absorbed dose and effective dose rates as well as cumulative doses. Improved modeling of the space environment is needed so that longer-term predictions may be of conditions in orbit.

For activities beyond low-earth orbit, there is also a need for improved forecasting capabilities [8] for SPE. There is also a need for development and validation of space radiation transport codes, accounting for neutrons, protons, light and heavy ions, mesons, and electromagnetic cascades. Radiation spectrometers are needed for combined measurements of neutron doses and doses from high-energy charged particles. As to biological effects, there is need for more sophisticated risk assessment methodology – at one extreme addressing late somatic and carcinogenic effects – at the other extreme addressing thresholds for symptoms affecting mission requirements, namely, central nervous system effects, dermal and immune issues.

Nuclear Power

Extending the operating lives of nuclear power plants brings on the need for increased attention to maintenance, corrosion control, and surveillance of piping and components. Similarly, extending the in-core operating life of individual fuel assemblies places intense demands on design, manufacturing, and quality assurance. Likewise, greater fuel "burnup" requires increased attention to actinide source inventories and secondary neutron production in spent fuels. Plant operating lives of 60 or more years need support of strong technical and manufacturing infrastructure.

No doubt the future will also see new generations of plants operating with radically advanced designs, with breeding capability, and using mixed ^{235}U/^{239}Pu fuels as well as fuels utilizing the ^{232}Th fuel cycle. These changes will bring on new design and operational challenges as well as the continued support of the physics community in broadening the evaluated nuclear data files (ENDF) and the evaluated nuclear structure data files (ENSDF).

Methods of storage and disposal of spent nuclear fuel continues to involve a complex mixture of technical, economic, political, and emotional issues. Resolution of the political and emotional issues seems to be the more demanding challenge.

Medical Applications

Challenges in nuclear medicine vary from nation to nation, but a recent survey [4] identified a number of universal concerns.

Hybrid imaging, employing dual use of CT, PET, MRI, and SPECT methods, introduces new combinations of radiation sources – positron emitters for PET, x-rays for CT, and gamma ray emitters for SPECT. Accounting for these mixed sources brings new challenges in facility design, treatment planning, patient and staff protection, as well as management of wastes. The same is true for radionuclides used in nuclear-medicine therapy such as delivery of radionuclides to malignant tumor cells using monoclonal antibody and related techniques.

In many countries, there is a need for improvement of domestic medical radionuclide production to alleviate the shortage of accelerator- and nuclear reactor–produced medical radionuclides available for research, diagnosis, and treatment. Finally, improvements in detector technology, image reconstruction algorithms, and advanced data processing techniques are needed to facilitate translation from research laboratory to the clinic.

Industrial and Commercial Activities

Sources of ionizing radiation find use in a very broad array of applications, examples being radiography and tracer techniques in manufacturing and density and moisture gauging in highway construction. Sources find their way into the home via ^{241}Am in smoke detectors and into public buildings via ^{3}H in exit signs. There is a very sad history of injury and death caused by the loss, abandonment, and theft of such sources. Better control of inventory, use, storage, and disposal of such sources is badly needed as well as better oversight by regulatory bodies.

Bibliography

Primary Literature

1. UN (1982) Sources and effects of ionizing radiation. Reports of the United Nations Scientific Committee on the effects of atomic radiation. United Nations, New York
2. UN (2000) Sources and effects of ionizing radiation. Reports of the United Nations Scientific Committee on the effects of atomic radiation. United Nations, New York
3. Eckerman KF, Ryman JC (1993) External exposure to radionuclides in air, water, and soil. Federal guidance report 12, EPA-402-R-93-081. U.S. Environmental Protection Agency, Washington, DC
4. National Research Council (2007) Advancing nuclear medicine through innovation. Committee on State of the Science of Nuclear Medicine. National Academy of Sciences, Washington, DC
5. World Nuclear Association (2006) Radioisotopes in industry (Information paper). www.world-nuclear.org/info/inf56.html
6. Fewell TR, Shuping RE, Hawkins KR (1981) Handbook of computed tomography and X-ray spectra. Report HHS (FDA) 81-8162
7. NCRP (2009) Ionizing radiation exposure of the population of the United States. Report no 160. National Council on Radiation Protection and Measurements, Bethesda

8. NCRP (2006) Information needed to make radiation protection recommendations for space missions beyond low-earth orbit. Report no 153. National Council on Radiation Protection and Measurements, Bethesda
9. NCRP (2000) Radiation protection guidance for activities in low-earth orbit. Report no 132. National Council on Radiation Protection and Measurements, Bethesda
10. NCRP (2002) Operational radiation safety program for astronauts in low-earth orbit. Report no 142. National Council on Radiation Protection and Measurements, Bethesda
11. NCRP (1987) Exposure of the population in the United States and Canada from natural background radiation. Report no 94. National Council on Radiation Protection and Measurements, Bethesda

Books and Reviews

Cohen BL (1986) A national survey of ^{222}Rn in U.S. homes and correlating factors. Health Phys 51:175–183

Cohen BL, Shah RS (1991) Radon levels in United States homes by states and counties. Health Phys 60:243–259

Eisenbud M (1987) Environmental radioactivity, 3rd edn. Academic, Orlando

Faw RE, Shultis JK (1999) Radiological assessment: sources and doses. American Nuclear Society, La Grange Park

Firestone RB, Shirley VS (eds) (1996) Table of isotopes, 8th edn. Wiley-Interscience, Malden

Haffner JW (1967) Radiation and shielding in space. Academic, New York

Glasstone S, Dolan PJ (eds) (1977) The effects of nuclear weapons. United States Departments of Energy and Defense, Washington, DC

ICRP (1987) Radiation dose to patients from radiopharmaceuticals. ICRP publication 53. Annals of the ICRP 18(1–4). International Commission on Radiological Protection, Pergamon Press, Oxford

Kocher DC (1981) Radioactive decay tables. Report DOE/TIC-11026. National Technical Information Service, Springfield

NAS (1971) Radioactivity in the marine environment. Report of the panel on radioactivity in the marine environment. Committee on Oceanography, National Research Council, National Academy of Sciences, Washington, DC

NAS (1988) Health risks of radon and other internally deposited alpha-emitters. Report of the BEIR Committee [The BEIR-IV Report]. National Research Council, National Academy of Sciences, Washington, DC

NAS (1990) Health effects of exposure to low levels of ionizing radiation. Report of the BEIR Committee [The BEIR-V Report]. National Research Council, National Academy of Sciences, Washington, DC

NCRP (1975) Natural background radiation in the United States. Report no 45. National Council on Radiation Protection and Measurements, Washington, DC

NCRP (1977) Radiation protection design guidelines for 0.1–100 MeV particle accelerator facilities. Report no 51. National Council on Radiation Protection and Measurements, Bethesda

NCRP (1984) Exposures from the Uranium Series with Emphasis on Radon and its Daughters. Report no 77. National Council on Radiation Protection and Measurements, Washington, DC

NCRP (1987a) Ionizing radiation exposure of the population of the United States. Report no 93. National Council on Radiation Protection and Measurements, Washington, DC

NCRP (1987b) Radiation exposure of the U.S. population from consumer products and miscellaneous sources. Report no 95. National Council on Radiation Protection and Measurements, Bethesda

NCRP (1989a) Exposure of the U.S. population from diagnostic medical radiation. Report no 100. National Council on Radiation Protection and Measurements, Bethesda

NCRP (1989b) Exposure of the U.S. population from occupational radiation. Report no 101. National Council on Radiation Protection and Measurements, Bethesda

NCRP (1989c) Radiation protection for medical and allied health personnel. Report no 105. National Council on Radiation Protection and Measurements, Bethesda

NCRP (2003) Radiation protection for particle accelerator facilities. Report no 144. National Council on Radiation Protection and Measurements, Bethesda

Shultis JK, Faw RE (2000) Radiation shielding. American Nuclear Society, La Grange Park

Slaback LA Jr, Birky B, Schleien B (1997) Handbook of health physics and radiological health. Lippincott Williams & Wilkins, Hagerstown

UN (1977/1982/1988/1993/2000) Report of the United Nations Scientific Committee on the effects of atomic radiation. United Nations, New York

Weber DA, Eckerman KF, Dillman LT, Ryman JC (1989) MIRD: radionuclide data and decay schemes. Society of Nuclear Medicine, New York

Radioactive Waste Management: Storage, Transport, Disposal

Audeen W. Fentiman
School of Nuclear Engineering, Purdue University,
West Lafayette, IN, USA

Article Outline

Glossary
Definition of the Subject
Introduction

High-level Radioactive Waste – Including Used (Spent) Nuclear Fuel (SNF)
Low-level Radioactive Waste
Transuranic Waste
Future Directions
Bibliography

Glossary

Fission Process by which a nucleus splits into two smaller nuclei, emitting two or three neutrons and energy.

Half-life The time required for half of the nuclei in a sample of a radioactive isotope to emit radiation and be transformed to another isotope.

High-level radioactive waste (HLW) Used nuclear fuel or the highly radioactive materials that are generated when used nuclear fuel is reprocessed.

Low-level radioactive waste (LLW) Radioactive waste that is not high-level radioactive waste, used nuclear fuel, or mill tailings.

Transuranic waste (TRU) Wastes that are not classified as high-level waste and contain more than 100 nCi/g of alpha-emitting transuranic isotopes ($Z > 92$) with half-lives of more than 20 years.

Used nuclear fuel or spent nuclear fuel (SNF) Terms used to designate nuclear fuel that has been irradiated in a reactor to produce power.

Definition of the Subject

Radioactive materials are widely used in our society, and when they are, radioactive wastes can be produced. In addition to being used to generate about 20% of the electricity used in the USA and 17% of the electricity used worldwide, radioactive materials are important in medicine, industry, and research. For example, radioactive materials are used to help diagnose and treat disease, as thickness gages in manufacturing, as components of some smoke detectors, to kill bacteria in food, to trace the movement of nutrients through plants, and to power spacecraft leaving the solar system. Just as there are many uses of radioactive materials, there are many types of radioactive waste, each of which must be stored for a time, treated (prepared either for disposal or recycling), transported, and ultimately disposed of in a licensed facility. Some radioactive materials are utilized in weapons production, an activity which also results in generation of radioactive wastes, although this chapter shall focus on civilian uses of radioactive materials. Proper management of radioactive wastes is essential to ensure that society continues to realize the benefits of radioactive materials without undue risk to human health or the environment, and all aspects of radioactive waste management are highly regulated.

Introduction

This chapter addresses four types of radioactive waste, used nuclear fuel, high-level radioactive waste (HLW), low-level radioactive waste (LLW), and transuranic waste (TRU). Used nuclear fuel (sometimes called spent nuclear fuel) is fuel that has reached the end of its useful life and has been taken out of a nuclear power plant. It is highly radioactive. High-level waste is a category of radioactive waste that includes used nuclear fuel and the highly radioactive wastes that are generated when used nuclear fuel is reprocessed. These two materials are expected to be disposed of in the same facility and thus are lumped together in one category. As the name implies, this waste is highly radioactive. Low-level radioactive waste is material that is not HLW or TRU; LLW typically consists of items with low concentrations of radioactive materials with relatively short half-lives (<100 years). There are several categories of low-level radioactive material, and the category assigned to any particular container of low-level waste depends on the type and amount of contamination on the waste material. Finally, transuranic waste is a very specialized type of waste containing materials that have an atomic number greater than that of uranium. One section of this chapter is devoted to each type of waste, and within each section, methods for storing, treating, transporting, and disposing of the waste will be discussed.

High-level Radioactive Waste – Including Used (Spent) Nuclear Fuel (SNF)

One hundred and four nuclear power plants in the USA generate about 20% of the electricity used each year [1]. Worldwide, over 435 nuclear power plants generate about 17% of the electricity [2]. Some countries rely heavily on nuclear power. For example, over 75% of the electricity in France is generated by nuclear power plants, and over 40% of the electricity in Sweden, Belgium, and Slovakia is from nuclear power [3].

The fuel for the nuclear power plants is uranium dioxide (UO_2), which is typically pressed into ceramic pellets. About 5% of the uranium in the fuel is ^{235}U which fissions, releasing energy that is ultimately converted to electricity. The other 95% of the uranium in the fuel is ^{238}U which is mostly inert although a small fraction of the ^{238}U atoms is converted to plutonium while the reactor is operating and another small fraction fissions. The ceramic UO_2 pellets are small cylinders about 0.6 in. long and 0.4 in. in diameter [4]. Stacks of pellets are sealed in zircaloy tubes about 12 ft long, and a 17 × 17 array of rods constitutes one typical fuel assembly. About 180 fuel assemblies are loaded into an average-sized pressurized water reactor. The other type of nuclear reactor commonly used in the USA, the boiling water reactor, is fueled with about 500–750 smaller fuel assemblies.

As the reactor operates, ^{235}U nuclei fission, after they absorb a neutron, breaking into two smaller nuclei. A small number of neutrons emitted during fission strike other ^{235}U atoms, causing more fissions, and releasing more energy. This process is known as a chain reaction. Eventually, most of the ^{235}U atoms in a fuel assembly have fissioned, and there are not enough of those atoms remaining to sustain the chain reaction. At that point, about one third of the fuel assemblies is removed from the reactor and fresh ones inserted in their place. Most nuclear power plants operate for 18–24 months before they need to be refueled. The fuel assemblies that are removed during refueling are referred to as used nuclear fuel or spent nuclear fuel.

The composition of the used nuclear fuel is shown in Table 1. Approximately 95% of the used nuclear fuel is ^{238}U. Less than 1% of the used fuel is plutonium. Some plutonium isotopes, primarily ^{239}Pu, fission and can be incorporated into new fuel rods. The reminder of the used fuel consists of fission products, that is, the atoms created when the ^{235}U atoms split, and minor actinides which were formed when some ^{238}U atoms absorbed neutrons. Many fission products are highly radioactive and emit penetrating gamma rays.

Radioactive Waste Management: Storage, Transport, Disposal. Table 1 Composition of used nuclear fuel [5]

Material	Percent of used fuel (%)
Uranium (^{235}U and ^{238}U)	95.6
Plutonium (all isotopes)	0.9
Minor actinides	0.1
Stable fission products	2.9
Radioactive fission products	0.5

When the used nuclear fuel is removed from the reactor, it is transferred to the spent fuel pool in a structure adjacent to the reactor building. The spent fuel pool is an in-ground, concrete, steel-reinforced, stainless steel-lined pool that is filled with at least 20 ft of water [6]. Used fuel assemblies are placed vertically in the pool. Water in the pool cools the fuel assemblies and serves to shield workers and equipment in the area of the pool from the radiation emitted by the fuel assemblies.

Perhaps surprisingly, the spent fuel pool at a nuclear power plant is not designed to be large enough to hold all of the used fuel that will be discharged from the nuclear power plant over its lifetime. A nuclear power plant's initial license is for 40 years. By early 2010, over half of the nuclear power plants in the USA had been granted license extensions of 20 years [7]. Thus, nuclear power plants are expected to operate at least 60 years, and it will be necessary to store some used fuel outside of the spent fuel pool. At the time the first nuclear power plants were built in the 1960s, it was presumed that used nuclear fuel would be reprocessed and recycled after the fuel assemblies had cooled for several years. Thus there was no need to build a spent fuel pool to hold all of the discharged fuel. In April 1977, US policy on reprocessing was changed when President Carter issued a statement that the USA would "defer indefinitely the commercial reprocessing and recycling of plutonium" [8]. The Nuclear Waste Policy Act of 1982 mandated construction of a geologic repository for permanent disposal of used nuclear fuel and required the Department of Energy to begin disposing of the used nuclear fuel not later than January 31, 1998 [9]. Once again, it did not appear that the spent fuel pool would need to hold all of the used nuclear fuel from a nuclear power plant, since there would be a place to send the used fuel for disposal. However, the geologic repository has not been built, and spent fuel pools at many of the older nuclear power plants are full or nearly full.

Many power plants are now moving some of their older used fuel assemblies that have been cooling in the spent fuel pool for many years to dry storage casks to make room in the pool for more fuel coming out of the reactor. The typical dry storage cask is a cylinder about 19 ft high and 8 ft in diameter made of concrete or steel. Dry storage casks come in different sizes, but they usually hold approximately 30 fuel assemblies and weigh about 100 t when loaded [10]. Spent fuel assemblies are sealed in an inner canister which is then placed in the dry storage cask. Some of the dry storage casks are designed to sit vertically on a concrete pad near the reactor building. In other dry storage systems, the casks are placed horizontally into a concrete bunker near the reactor building.

Across the USA, as of 2010, there are about 60,000 t [11] of used nuclear fuel stored either in spent fuel pools or dry storage casks at the nuclear power plants where it was generated. (Because uranium is a very dense material (19.05 g/cm^3), denser even than lead (11.35 g/cm^3), a ton of used fuel does not occupy much space. If all of the used fuel assemblies that have been discharged from US nuclear power plants since they began to operate were stacked on a football field 100 yards long and 53⅓ yards wide, the stack would be about 6½ yards high.) The Nuclear Regulatory Commission which is responsible for overseeing safety at the nuclear power plants has said that the used fuel can be safely stored at the power plant where it was generated for up to 100 years [12].

Ultimately, the used nuclear fuel must be disposed of permanently or reprocessed with as many of the constituents as possible being recycled and reused and the remainder being disposed of permanently. Until a decision about the final disposition of used nuclear fuel is made, the used fuel will continue to be stored either at the power plants where it was generated or at central storage facilities. Regardless of where the fuel is stored or whether it is reprocessed or buried, a system for transporting the used nuclear fuel will be required. Regulations governing the transportation system have been in place for decades, and some used fuel has been moved in licensed transportation casks. The next several paragraphs will describe recent US policy for disposing of used nuclear fuel, the options for central storage of used fuel if the fuel currently in dry storage casks is moved from the reactors to a central location, and the transportation system in place to move that used fuel.

Current US policy for disposal of used nuclear fuel was set by the Nuclear Waste Policy Act of 1982 and its 1987 amendments which provided for permanent disposal of used nuclear fuel in a deep geologic repository. The repository was to hold 70,000 t of used nuclear fuel and vitrified high-level waste from Department of Energy (DOE) facilities. The DOE was responsible for designing, building, and operating the repository, and the DOE was to begin taking used fuel from nuclear power plants by January 31, 1998. The Office of Civilian Radioactive Waste Management was established within DOE to manage both the repository and the transportation system for used nuclear fuel. In addition, the law required a monitored retrievable storage facility (MRS) where used fuel from reactors throughout the USA could be stored temporarily and then put into standard packages for disposal. The bill also established the Nuclear Waste Fund to pay for the design, construction, and operation of the used nuclear fuel management system. A charge of one mill (one tenth of a cent) for every kilowatt hour of electricity generated by a nuclear power plant is the source of money for the Nuclear Waste Fund. Between 1982 and 2007, $27.3 billion dollars were collected for the Nuclear Waste Fund [13].

Three possible sites for the deep geologic repository were identified, and each was to be characterized to determine whether it was an appropriate location for nuclear waste disposal. The sites were in (1) Deaf Smith County, Texas, (2) Richland, Washington, on the Hanford Site, and (3) Yucca Mountain, Nevada, near the Nevada Test Site where nuclear weapons had been tested underground. Each of the three sites offered a different type of rock in which the repository would be located. The Texas site was in salt, the Washington site in basalt, and the Nevada site in tuff. Local opposition to construction of a nuclear waste repository was strong in all three locations, and efforts to characterize the sites were often thwarted. Little progress was made toward identifying the best location for the repository.

In 1987, the US Congress passed the Nuclear Waste Policy Amendments Act. This law identified the Yucca Mountain site in the Nevada desert as the location of the nation's deep geologic repository. The law established a position of Nuclear Waste Negotiator to find a volunteer site for the MRS since the site in Tennessee identified pursuant to the 1982 Nuclear Waste Policy Act was rejected.

The geologic repository for used nuclear fuel was required to be licensed by the Nuclear Regulatory Commission (NRC). Following passage of the 1987 law, work began at the Nevada site to gather scientific evidence required to prepare an application for a license to construct and operate the geologic repository. Information required for the license application is specified in 10 CFR Part 60, Disposal of High-Level Radioactive Wastes in Geologic Repositories [14]. Scientists working on the application needed to show that they understood not only the characteristics of each of the components of the repository, but also how the components would interact and perform over time. Since the repository was to accommodate both used nuclear fuel from commercial nuclear power plants and vitrified high-level waste from nuclear weapons programs, data on composition of both waste forms as well as chemical and physical properties were required. Information required on packages that confined the waste included proposed materials, dimensions, and response to heat, pressure, radiation, water, and other possible corrosive chemicals in the rock and soil. Detailed information on the design of the engineered repository was required. DOE had to show that the repository would be safe for workers while the waste was being emplaced, that any specific container of buried waste could be retrieved for 50 years after it had been emplaced, and that the repository would confine the waste over a specified period. Data required for the site, itself, included characteristics of the rocks and soil, groundwater speed and direction, and an inventory of flora and fauna species in the area of the repository.

Between 1987 and 2007, the team of scientists characterizing Yucca Mountain spent approximately $7 billion gathering data on the site and preparing the required documents including an environmental impact statement and the license application. The repository design called for a network of tunnels approximately 1,200 ft below the surface. Each tunnel was to be reinforced with steel supports to keep the rock from collapsing onto the waste packages. Rails on the floor of each tunnel would allow casks to be moved into place. Casks designed to hold about 30 used fuel assemblies or half a dozen cylinders of vitrified waste from DOE facilities would be made of thick steel. An inverted U-shaped titanium shield was proposed to cover the casks in each tunnel to divert any water that might reach the repository from the desert above.

On June 3, 2008, DOE delivered to the Nuclear Regulatory Commission the application for a license to construct and operate the nuclear waste repository at Yucca Mountain. According to the Nuclear Waste Policy Act, the NRC had 3 years to review the application but could request an additional year. Following the presidential election in 2008, funding for the Yucca Mountain Project was reduced, and March 2010, the DOE withdrew its application for a license for a nuclear waste repository at Yucca Mountain "with prejudice." At almost the same time, the DOE created the Blue Ribbon Commission on America's Nuclear Future. The charter of this Commission was "to conduct a comprehensive review of policies for managing the back end of the nuclear fuel cycle, including all alternatives for the storage, processing, and disposal of civilian and defense used nuclear fuel, high-level waste, and materials derived from nuclear activities." The Commission is to complete its report in 24 months.

The Blue Ribbon Commission is likely to consider both direct disposal of used nuclear fuel assemblies in a deep geologic repository and reprocessing of used nuclear fuel followed by recycling of many of the components and utilization, treatment, or disposal of the remaining materials. Direct disposal in a geologic repository was discussed earlier. Reprocessing of used nuclear fuel is briefly addressed here.

Reprocessing of used nuclear fuel currently involves chopping up the used fuel rods, dissolving the fuel in a concentrated nitric acid solution, and chemically separating the various elements in the used fuel. The uranium and plutonium, which constitute about 96% of the used fuel, can be recycled and used in fabricating new fuel rods. The remaining material is typically dried, mixed with glass frit, melted, and poured into metal cylinders. Many countries that rely heavily on nuclear power either reprocess their own used nuclear fuel or send it to other countries that have reprocessing facilities. The USA does not reprocess used fuel from commercial nuclear power plants.

The reprocessing method currently used in most countries is the PUREX method which was developed in the USA to reprocess fuel from government-run reactors to recover plutonium for use in nuclear weapons.

A small amount of used fuel from commercial nuclear power plants was reprocessed at a facility in West Valley, New York, between 1966 and 1972, but the facility was shut down because it was not economical. Since the USA decided not to reprocess used nuclear fuel in 1977, no additional reprocessing of commercial nuclear fuel has occurred. One objection to use of the PUREX method was that it isolates plutonium from other materials in the used fuel, supposedly making it easier for terrorists to divert the plutonium.

Research is being conducted in the USA on reprocessing methods that would not isolate plutonium. Some of the methods under consideration are for reprocessing used fuel from the light water reactors that are currently operating in the USA. Other methods are being developed for different types of fuel that might be used in the next generation of reactors (commonly referred to as Generation IV) now being designed for use around the world. Reprocessing facilities are complex and expensive to build. Japan completed construction of a nuclear fuel recycling facility in Rokkasho at a cost of approximately $20 billion [15]. It is likely that US policy makers will want more information on the types of nuclear power plants that will be operating in the USA over the next century and results of research programs on reprocessing methods in hand before deciding what type of reprocessing facility, if any, to build in the USA.

Since it appears that the USA will not be disposing of used nuclear fuel or reprocessing it in the near future, storage of used nuclear fuel at a central facility (or facilities) is being studied. Central storage is not a new concept. The Nuclear Waste Policy Act of 1982 called for establishing a Monitored Retrievable Storage (MRS) facility, and the Nuclear Waste Policy Amendments Act of 1987 provided for a person to seek a community willing to host the MRS. No MRS has been sited. A consortium of utilities, called Private Fuel Storage, negotiated with the Skull Valley Band of the Goshute Indian Tribe to establish a central used fuel storage facility on the Tribe's land in Utah. An application was submitted on the NRC for a license to construct and operate the storage facility, and on February 21, 2006, the NRC granted a license, but said that construction could not begin until it obtained "necessary approvals from other agencies, including the Bureau of Land Management, the Bureau of Indian Affairs, and the Surface Transportation Board" [16]. Neither the Bureau of Land Management nor the Bureau of Indian Affairs has approved the site. At the request of Senators Barbara Boxer, Harry Reid, and John Ensign, the US General Accountability Office did a study of a central storage facility option and an on-site storage facility option, along with the Yucca Mountain repository. The study, which was issued in November 2009, concluded that while a centralized interim storage facility could be built relatively quickly, finding a site could be difficult, and since the facility would not be a final disposal site, any waste going to the centralized storage facility would have to be transported twice [17]. The Blue Ribbon Commission on America's Nuclear Future is also likely to consider options for a central storage facility.

Eventually, the used nuclear fuel currently stored at the nuclear power plants must be transported to a central storage facility, a processing facility, or a permanent disposal site. Some used nuclear fuel has been transported in the USA, for example, between nuclear power plants owned by the same company or from a power plant to a government research facility. Both the NRC and the US Department of Transportation have regulations governing transportation of used nuclear fuel. The NRC regulation is Title 10, Part 71 of the Code of Federal Regulations (10 CFR Part 71) [18], which specifies requirements for the packages that carry the highly radioactive material, called Type B packages. The regulation also specifies how packages for transporting used nuclear fuel are to be approved. Each Type B package must have a Radioactive Material Package Certificate of Compliance from the NRC. Procedures for applying for the Certificate of Compliance can be found in NUREG-1617, "Standard Review Plan for Transportation Packages for Spent Nuclear Fuel" [24]. Packages for used nuclear fuel are designed for transportation by truck and by rail.

To receive a Certificate of Compliance, a Type B package, which is often called a shipping cask, must undergo a series of tests that simulate accident conditions a cask might encounter en route. The first four tests, conducted sequentially on a single cask, are:

1. Drop test. Drop the cask from 30 ft onto a hard, unyielding surface in an orientation most likely to damage the cask.

2. Puncture test. Drop the cask from 40 in. onto a 6-in. diameter shaft in an orientation most likely to result in damage.
3. Fire test. Engulf the cask fully in a fire at least 1,475°F for 30 min.
4. Immersion test. Place the cask under 3 ft of water for 30 min [19].

In order to receive the Certificate of Compliance, the cask must not release any radioactive materials during or following the series of tests. A new, undamaged cask must pass a fifth test, immersion in water at a pressure equivalent to that exerted by water 50 ft deep, before the Certificate can be issued.

Several different shipping casks have been designed to be transported by truck. Those casks carry between 1 and 9 used fuel assemblies. A typical rail cask carries 36 fuel assemblies, but the number varies because there are different sizes of fuel assemblies. The shipping casks are usually cylindrical with their walls made of several layers of different materials. The innermost and outermost layers are usually steel to provide structural strength. One layer of material between the steel shells is designed to absorb gamma rays emitted by the used fuel. Steel, lead, and depleted uranium are some materials that are used to absorb gamma rays. Another layer is made of a material that slows down and absorbs neutrons. *The Radioactive Materials Packaging Handbook*, [20] written at the Oak Ridge National Laboratory in 1998 contains detailed information required to design and manufacture a shipping cask for used nuclear fuel.

NRC regulations govern routes to be used for transporting used nuclear fuel and physical protection for the shipments. The Department of Transportation regulations specify methods for selecting routes (49 CFR 397) and labeling of the shipping casks (49 CFR 172). In addition, drivers of trucks transporting radioactive materials must meet training and experience requirements and undergo a background investigation.

Low-level Radioactive Waste

Low-level radioactive waste is defined in the Low Level Radioactive Waste Policy Act of 1980 as radioactive waste that is not high-level radioactive waste, spent nuclear fuel, or mill tailings [21]. There are four classes of low-level radioactive waste (LLW) based on the concentration of radioisotopes with short half-lives and the concentration of radioisotopes with longer half-lives in the waste. The classes are designated as A, B, C, and Greater Than Class C (GTCC). Tables 1 and 2 in 10 CFR Part 61 are used to determine the class of LLW in a particular container.

LLW is generated by almost any activity that involves radioactive material. Some examples of LLW are lab coats and shoe covers worn when working with radioactive materials, medical equipment involved in treating a patient with radiopharmaceuticals, and laboratory supplies and equipment used in experiments involving radioactive tracers. Radioactive tracers are very commonly used. They are radioactive isotopes that move through a system (e.g., the human body, a growing plant, or an ecosystem) along with the material being studied. Since small amounts of radiation can be detected, the researchers can measure radiation from the tracer to determine how the substance of interest is moving through the system. For example, tracers are used in the development of virtually all new medicines to study how the medicine or its metabolites move through the body. Hospitals, universities, and research laboratories operated by corporations or government agencies often generate LLW.

The largest amount of commercial LLW is generated by industry, including the nuclear power industry. Equipment and materials from power plant maintenance activities, samples collected during environmental monitoring, and protective clothing are disposed of as LLW. In addition, LLW is generated during the various stages of nuclear fuel fabrication.

Often LLW is initially in liquid form. For example, it is sometimes more cost-effective to wash protective clothing than to dispose of it. Likewise, it may be more economical to decontaminate some equipment using a solvent than to dispose of the equipment. In these cases, the wash water and the solvent become LLW. Environmental samples are routinely dissolved in a liquid for analysis, yielding liquid LLW. Since LLW disposal facilities will not accept large amounts of liquid wastes, most liquid LLW must be solidified. Several methods can be used including evaporation, ion exchange, flocculation, and filtration. The liquid waste could simply be mixed with a solidifying agent such as concrete. However, since the cost of disposal is determined, in part, by the volume of the waste, methods that minimize the volume are usually preferred.

Dry LLW is typically treated to reduce its volume prior to packaging for shipment to a disposal facility. Compaction, which results in a volume reduction to one half or one third of the initial volume, or supercompaction, which can result in a volume reduction to one tenth of the original volume are commonplace. Incineration can reduce the volume to 1% of the original and is especially useful for combustible materials like wood that cannot be compacted easily.

Three disposal facilities for commercial LLW are currently operating in the USA. They are in Barnwell, South Carolina, Richland, Washington, and Clive, Utah. Another one has been proposed in Texas. LLW disposal sites have been available since the early 1960s. Four early sites have been closed. They were in Sheffield, Illinois, Maxey Flats, Kentucky, West Valley, New York, and Beatty, Nevada.

While LLW disposal facilities have been operating in the USA since the 1960s, the law which currently governs LLW disposal was not passed by the US Congress until 1980. The Low-Level Radioactive Policy Act of 1980 made each state responsible for arranging for disposal of its own LLW. States were encouraged to form compacts, groups of states that could collaborate to build one LLW disposal facility to serve all of the compact's members. If a state within a compact built a LLW disposal facility, that facility would not be required to accept LLW from any state outside of the compact. States that did not belong to a compact and chose to build their own disposal facility would not be able to exclude waste from other states. The law required states to form compacts by 1986. States did not meet that deadline, and in 1985, the Low Level Radioactive Waste Policy Amendments Act was passed. It gave the states until 1992 to form compacts. At one time, all states have belonged to a compact. Some have left the compact, and some compacts have been reconstituted. However, no compact has built a new LLW disposal facility.

Until 2008, the lack of new disposal facilities did not impact LLW generators' ability to dispose of their waste. The Richland, Washington, facility accepted Class A, B, and C waste from the Northwest and Rocky Mountain compacts, and the Barnwell, South Carolina, facility accepted Class A, B, and C waste from the rest of the country. The Clive, Utah, facility accepted only Class A waste, but it had applied for a revised license to accept Class B and C waste. The Barnwell facility announced that after July 1, 2008, it would not accept Class B and C waste from states outside of the Atlantic Compact. As of early 2010, the Clive, Utah, facility did not yet have a license to accept Class B and C waste. Thus, as of July 2008, the 36 states that had previously been served by the Barnwell facility have had no place to send their Class B and C wastes. Since nearly 99% of commercial LLW is Class A waste and can be sent to a disposal facility, most generators have only small amounts of Class B and C waste and will have space to store that waste until a new facility can be built or some other solution can be found [22].

Packaging requirements are specified for each class of LLW in 10 CFR Title 61.56. Some requirements for Class A waste are: no cardboard or fiberboard containers may be used, wastes cannot be explosive or pyrophoric, emit toxic gases, or contain biological pathogens or infectious material and if the package contains liquids, it must also have enough absorbent material to absorb twice the volume of liquid present. Packing for Class B waste must meet all of the requirements for Class A waste packaging plus a stability criterion which ensures that packages will remain intact when other packages are stacked on top of them. Class C waste packaging must meet all of the requirements of Class B waste packaging plus provide a barrier to inadvertent intrusion.

Low-level radioactive waste is typically disposed of in shallow land burial facilities. A facility must be designed to minimize contact of the buried waste with water. A trench about 30 ft deep and 100 ft wide is excavated in soil that drains well. Containers are placed in the trench in a way that minimizes void space, and any voids that are created are filled with sand. When a portion of the trench is full, it is covered with a cap that is designed to divert any water that falls on it away from the waste.

Transuranic Waste

Transuranic wastes are defined in 40 CFR 191, Environmental Standards for the Management and Disposal of Spent Nuclear fuel, High Level and Transuranic Waste, as wastes that are not classified as high-level waste but contain more than 100 nCi/g of alpha-emitting transuranic isotopes (materials with $Z > 92$) with half-lives of more than 20 years. Transuranic isotopes are generated in a reactor when U-238 absorbs neutrons. Most transuranic waste in the USA has been generated during weapons production when used nuclear fuel from government-operated reactors was reprocessed to recover plutonium for weapons. If the USA begins to reprocess used nuclear fuel from commercial nuclear power plants, transuranic wastes will be generated at the commercial fuel reprocessing facilities as well.

Most transuranic (TRU) waste is solid. Typical wastes are protective clothing or equipment that has been contaminated with transuranic isotopes. TRU waste is packaged in steel, concrete, or wooden boxes. Since most of the transuranic isotopes emit alpha particles which cannot penetrate a sheet of paper, most TRU waste (about 97%) is referred to as Contact Handled TRU, meaning that people can handle the packages. Only 3% of the TRU must be handled remotely [23].

Transuranic waste is disposed of in the Waste Isolation Pilot Plant (WIPP). WIPP is a deep geologic repository in salt located near Carlsbad, New Mexico. It has been accepting and disposing of TRU waste since 1999. The US Department of Energy Report DOE-WIPP-069, Waste Acceptance Criteria for the Waste Isolation Pilot Plant, outlines requirements for TRU waste shipped to the facility for disposal. They include specifications for the container, data to accompany the package, and radiologic, physical, chemical, and gas generation properties of the waste.

TRU waste is transported to WIPP in Type B packages which are certified by the Nuclear Regulatory Commission after undergoing the rigorous tests described in the section on used nuclear fuel.

Future Directions

Future directions for radioactive waste management in the USA vary depending on the type of radioactive waste. Transuranic wastes are likely to continue to be shipped to the Waste Isolation Pilot Plant (WIPP) for disposal. For decades, low-level radioactive wastes have been disposed of in shallow land burial facilities. Currently, several states do not have a location for disposal of Class B and C wastes, but those constitute only about 1% of the low-level wastes and can be stored at the generation site until a disposal facility is available. Efforts to open new or expand existing low-level waste disposal facilities are ongoing.

High-level radioactive waste disposal policy in the USA is being reviewed. The Blue Ribbon Commission on America's Nuclear Future has been established (2010) to consider the alternatives. Meanwhile, used nuclear fuel is being stored at reactor sites and can be kept there for several decades while the federal government adopts a policy for its disposal and constructs the facilities required to carry out the policy. The nation has experience with all of the components of the system required to treat and dispose of HWL. Transportation casks for HLW have been built, certified, and used to transport that material. Used nuclear fuel belonging to the federal government and some used commercial fuel have been reprocessed. Research on advanced reprocessing methods is being conducted at national laboratories and universities. Research done at Yucca Mountain and experience at WIPP have provided extensive data on deep geologic repositories, which are the type of facility in which used nuclear fuel or reprocessing waste is likely to be placed. Several other countries that rely heavily on nuclear power for their electricity are currently reprocessing used nuclear fuel and conducting research on geologic repositories. Technical information related to treatment, storage, transportation, and disposal of high-level radioactive waste will be available when the policy decisions are made.

Bibliography

Primary Literature

1. US Energy Information Administration (2009) United States percent of electricity from nuclear. http://www.eia.doe.gov/cneaf/electricity/epm/epm_sum.html. Accessed 15 Mar 2010
2. World Nuclear Association (2008) World nuclear power reactors & uranium requirements. http://www.world-nuclear.org/info/reactors.html. Accessed 15 Mar 2010
3. World Nuclear Association (2008) Nuclear shares of electricity generation. http://www.world-nuclear.org/info/nshare.html. Accessed 15 Mar 2010
4. Murray RL (1988) Nuclear energy. Pergamon Press, Oxford
5. Ryskamp JM (2003) Nuclear fuel cycle. http://nuclear.inel.gov/docs/papers-presentations/nucle_fuel_CYCLE_3-5-03.PDF. Accessed Jun 2008

6. US Nuclear Regulatory Commission (2007) Spent fuel pools. http://www.nrc.gov/waste/spent-fuel-storage/pools.html. Accessed Mar 2010
7. US Nuclear Regulatory Commission (2010) Nuclear power plant license extensions. http://www.nrc.gov/reactors/operating/licensing/renewal/applications.html#completed. Accessed Mar 2010
8. PBS Frontline (2010) Carter nuclear fuel reprocessing. http://www.pbs.org/wgbh/pages/frontline/shows/reaction/readings/keeny.html. Accessed Mar 2010
9. Nuclear Waste Policy Act of 1982 as amended (2010) http://www.nrc.gov/reading-rm/doc-collections/nuregs/staff/sr0980/v2/sr0980v2.pdf#pagemode=bookmarks&page=141. Accessed Apr 2010
10. Board on Radioactive Waste Management (2006) Safety and security of commercial spent nuclear fuel storage: public report. National Academies Press, Washington
11. How many tons of spent nuclear fuel on hand – in what year http://epw.senate.gov/public/index.cfm?FuseAction=Minority.PressReleases&ContentRecord_id=f2cbe309-802a-23ad-4925-643845f220b5&Region_id=&Issue_id
12. Waste Confidence Decision, 49 FR 34694, August 31, 1984, as amended at 55 FR 38474, September 18, 1990; 72 FR 49509, August 28, 2007, Federal Register
13. Congressional Budget Office Testimony on the Federal Government's Liabilities under the Nuclear Waste Policy Act (2007) Nuclear waste fund balance. http://www.cbo.gov/doc.cfm?index=8675&type=0. Accessed Mar 2010
14. US Nuclear Regulatory Commission (1981) 10 CFR Part 60, disposal of high-level radioactive wastes in geologic repositories
15. vonHippel F (2009) Why reprocessing persists in some counties and not in others: the costs and benefits of reprocessing – for the Non-Proliferation Education Center. www.npec-web.org/.../vonhippel%20-%20TheCostsandBenefits.pdf. Accessed Mar 2010
16. US Nuclear Regulatory Commission press release No. 06-028 (2006) Private Fuel Storage license application. http://www.nrc.gov/reading-rm/doc-collections/news/2006/06-028.html. Accessed Mar 2010
17. US General Accountability Office (2009) Nuclear waste management: key attributes, challenges, and costs for the Yucca mountain repository and two potential alternatives. GAO-10-48
18. US Code of Federal Regulations (CFR) (2010) 10 CFR Part 71 http://www.nrc.gov/reading-rm/doc-collections/cfr/part071. Accessed Mar 2010
19. Kutz M (2009) Environmentally conscious materials handling. Wiley, Hoboken
20. Arnold ED, Shappert LB, Bowman SM (1998) Radioactive materials packaging handbook: design, operations, and maintenance. Department of Energy, Oak Ridge National Laboratory, Oak Ridge
21. Low-level Radioactive Waste Policy Amendments Act of 1985. http://www.nrc.gov/reading-rm/doc-collections/nuregs/staff/sr0980/v2/sr0980v2.pdf#pagemode=bookmarks&page=15. Accessed Apr 2010
22. US General Accounting Office (2004) Low-level radioactive waste: disposal availability adequate in the short term but oversight needed to identify any future shortfalls. GAO-04-604
23. Saling JH, Fentiman AW (2002) Radioactive waste management. Taylor & Francis, New York
24. US Nuclear Regulatory Commission (2000) NUREG-1617 Standard Review Plan for Transportation Packages for Spent Nuclear Fuel. http://www.nrc.gov/reading-rm/doc-collections/nuregs/staff/sr1617/. Accessed Mar 2010

Books and Reviews

Office of Civilian Radioactive Waste Management Website (2010). http://www.ocrwm.doe.gov/. Accessed Mar 2010

Ohio State University Extension (2010) Low-level radioactive waste fact sheets. http://ohioline.osu.edu/rer-fact/index.html. Accessed Mar 2010

US Department of Energy (2002) Final environmental impact statement for a geologic repository for the disposal of spent nuclear fuel and high-level radioactive waste at Yucca mountain, Nye County, Nevada. DOE/EIS-0250

US General Accountability Office (2008) Low-level radioactive waste: status of disposal availability in the United States and other countries. GAO-08-813T

US General Accountability Office (2007) Low-level radioactive waste management: approaches used by foreign countries may provide useful lessons for managing US Radioactive Waste. GAO-07-221

US Nuclear Regulatory Commission Website, Radioactive Waste (2010). http://www.nrc.gov/waste.html. Accessed Mar 2010

Waste Isolation Pilot Plant Website (2010). http://www.wipp.energy.gov/index.htm. Accessed Mar 2010

Radionuclide Fate and Transport in Terrestrial Environments

John C. Seaman[1], Kimberly A. Roberts[2]
[1]Environmental Geochemistry, Savannah River Ecology Laboratory, The University of Georgia, Aiken, SC, USA
[2]Radiological Performance Assessment, Savannah River National Laboratory, Aiken, SC, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Radiochemistry Background

Radionuclide Sources
Soils as the Repository for Radioactive Waste
Important Soil Processes
Models for Describing Radionuclide Sorption and Plant Uptake
Characteristics of Specific Radionuclides
Future Directions
Acknowledgments
Abbreviations
Bibliography

Glossary

Actinides Row of elements in the periodic table and so-called 5f elements with atomic number 89 (actinium) through 103 (lawrencium).

Amphoteric The ability to act as either an acid or a base depending on conditions.

Biological half-life ($t_{\frac{1}{2}}^{bio}$) A measure of the rate at which radionuclides are excreted from an organism defined as the time required for the twofold decrease in activity resulting from processes other than radioactive decay.

Ecological half-life ($t_{\frac{1}{2}}^{eco}$) A measure of the rate at which the exposure and bioavailability of a radionuclide decreases with time due to processes other than radioactive decay.

Lanthanides Fourteen elements with atomic numbers 58 through 71 generally forming trivalent cations.

Radioactive half-life ($t_{1/2}$) The time required for the activity of an isotope to be reduced by half.

Transuranics Elements with atomic number 92 (uranium) and greater.

Definition of the Subject

The soil environment becomes the ultimate repository for radioactive waste in terrestrial systems, whether intentional or not. Therefore, a fundamental understanding of radionuclide migration and cycling in the soil environment is critical to the development of environmental policies that are protective of both human health and the environment. Radionuclides may be assimilated and/or immobilized into the organic matter pool within the soil microbiota and absorbed by the roots of higher order plants as the first steps in trophic transfer in terrestrial systems. An added complication in dealing with radionuclides is that decay and fission processes result in radioactive products with significantly different chemical properties from the parent material. The primary objective of this chapter is to describe the physical and chemical processes that impact the fate and transport of radionuclides in the soil environment, including a discussion of the results from recent studies and the identification of appropriate reference material for more in-depth analysis. For the current discussion, the term "soil" will be used in a generic sense to represent unconsolidated geologic material, including both vadose (unsaturated) and saturated groundwater systems, as the underlying processes are the same.

Introduction

In addition to nuclear weapons testing and accidental releases, natural and anthropogenic radioactive substances may enter the terrestrial environment from natural uranium (U) ore deposits, U mining and processing, waste discharges and disposal efforts. Thus, various soil processes as impacted by the chemical speciation of the radionuclide of interest control both ecosystem exposure (i.e., wildlife, human, etc.) and the success of any waste containment or remediation effort. In the USA, remediation efforts have generally focused on limiting radiation exposure to humans under the assumption that such an approach is also protective of the general environment, while the European community has applied more of an ecosystem approach to environmental policy [1, 2].

In recent years the interest in nuclear energy has been renewed due to concerns over the rise of greenhouse gasses and their impact on global climate change, the cost and availability of fossil fuel sources, and political instability and conflict within oil exporting countries. Since 2006, the US Nuclear Regulatory Commission (NRC) has received preliminary licensing applications for the construction of more than 30 new reactors. However, acceptance of nuclear power in the USA in particular is still limited by the public's perception of adverse safety, environmental, and health effects associated with nuclear facilities [3].

Despite improvements in reactor design, several important environmental issues face the nuclear power industry. Proposed expansion could exacerbate environmental concerns associated with U mining and processing [4]. Furthermore, the disposition of spent nuclear fuel (SNF) remains a major impediment to industry expansion in the USA with the future of the Yucca Mountain repository still in question. Currently, 54,000 metric tons of SNF are stored on site at existing nuclear facilities in the USA, with plans to require new facilities to provide enough on-site storage to accommodate the entire inventory of SNF generated during the 60-year lifetime of current reactor designs [5, 6]. Furthermore, the growing inventory of SNF at existing power stations in the USA will likely exceed the legislatively defined capacity for the Yucca Mountain repository within the next decade or so, even without expansion of the nuclear industry [7]. The eventual disposition of such waste remains unclear as debate continues regarding the cost and technical limitations associated with long-term storage as opposed to various SNF reprocessing technologies that increase the risk of nuclear proliferation [5, 6, 8].

Radiochemistry Background

Before further discussion concerning the fate of radionuclides in soils, a limited review of radiochemistry is warranted. Interested readers seeking a more thorough discussion are directed to one of several comprehensive texts that are currently available, i.e., [9–11]. Isotopes of an element differ by mass but display the same chemical properties, except for minor mass-dependent differences. Most readers are familiar with the conventions for distinguishing various isotopes of a given element:

$$ {}^{A}_{Z}X \tag{1}$$

where the mass number A represents the sum of protons (Z) and neutrons (N) within the nucleus of the element in question, X, with Z often omitted because it is defined by the element symbol [12, 13].

Radioactive decay is a spontaneous, irreversible process that releases energy, the magnitude of which depends on the element and mode of decay, as unstable isotopes transform to a lower energy state. Unlike other chemical reactions/transformations, radioactive decay is insensitive to environmental conditions, like temperature and pressure. When an isotope is radioactive, it is referred to as a radionuclide. The standard SI unit for radioactivity is the Becquerel (Bq), which is equivalent to one disintegration per second. For environmental samples, radioactivity is often reported in terms of picocuries (1 pCi = 10^{-12} Curies), with 1 pCi equivalent to 0.037 Bq.

Each radionuclide decays at a constant rate. The decay constant, λ, is the probability of decay per unit time and is related to half-life ($t_{1/2}$), the time required for the activity of an isotope to be reduced by half. For a more detailed discussion on the mathematics of radioactive decay, the reader is referred to Ivanovich [14]. In an ecological context, the biological half-life ($t_{\frac{1}{2}}^{bio}$) is a measure of the rate at which the radioactivity within an organism decreases by processes other than decay after the exposure source has been removed. Furthermore, the ecological half-life ($t_{\frac{1}{2}}^{eco}$) refers to the processes of redistribution and "aging" in the soil environment that control exposure levels for a given organism.

Another useful term in discussing radionuclides is specific activity which is the number of decays per unit time per amount of substance (usually mass). The shorter the half-life, the higher the specific activity and the more radioactive an isotope is considered. The concept of specific activity is important in terms of assessing radioactivity in the environment. A radionuclide with a long half-life may not be a concern, but could have short-lived daughters and thus considered to be more highly radioactive.

Alpha (α), beta (β), and gamma (γ) are the most common forms of radioactive decay. Other decay modes such as electron capture and proton and neutron emission are far less common, and beyond the current general discussion. Alpha decay is the emission of helium ions (${}^{4}_{2}He$) and generally observed for elements heavier than lead (Pb) on the periodic table, most actinides and some lanthanide elements. For example:

$$ {}^{A}_{Z}X \rightarrow {}^{A-2}_{Z-4}X +{}^{4}_{2}He \tag{2}$$

with α particles generally having kinetic energy in the range of 3–9 MeV (mega-electron volts; 10^{6} eV), but displaying very limited ability to penetrate matter because of their large mass and charge. A sheet of paper or skin is all that is needed to prevent penetration of α particles. Thus, the big health threat from α decay is through inhalation or ingestion, as the particles trapped inside the body can damage sensitive tissue.

Although derived from the nucleus, β particles are essentially electrons with a mass of 5.49×10^{-4} amu (atomic mass units), and a charge of -1. For isotopes with an unstable N/Z ratio, β decay is a spontaneous exothermic process where

$$^{A}_{Z}X \rightarrow {}_{z+1}{}^{A}X + \beta^{-} + \text{decay energy} \tag{3}$$

resulting in an increase in Z with no change in A. Due to the limited energy release, β decay is commonly measured by liquid scintillation counting. In terms of penetration, β particles are intermediate between α particles and γ rays. Plexiglas is typically sufficient shielding for these particles. Tritium (^{3}H) is an example of a β emitter.

The two decay processes discussed above, i.e., α and β decay, may leave the nucleus in its ground state, or more frequently in an excited state. The excited nucleus may then release energy through the emission of electromagnetic radiation known as γ radiation in returning to the ground state. Although somewhat similar to x-rays, γ rays display a very high frequency (~10^{19} Hz) and a shorter wavelength than x-rays, with energies ranging from a few thousand keV to 10 MeV. Gamma rays are produced from transitions within the nucleus, while x-rays are emitted from the electron shell. Two additional processes, internal conversion and de-excitation are means by which the excited nucleus may return to the ground state. All three processes are known as γ transitions because no change in A or Z is observed. In an example of γ decay

$$^{A}X^{*} \rightarrow {}^{A}X + \gamma \tag{4}$$

X^{*} represents a nucleus in an excited state. Gamma decay can occur between multiple lower energy states before reaching the ground state. Since γ rays carry no charge, their interaction with matter is limited and depends on the density of the absorber, with much greater penetration than observed for α or β radiation [10, 15]. Lead (Pb) is commonly used as a shield for γ rays.

Fission Products

The isotopes ^{235}U and ^{239}Pu have the unique ability to split into two atoms (i.e., fission products) when impacted by a slow moving neutron, releasing additional neutrons plus a significant amount of energy:

$$\begin{aligned} {}^{235}U + {}^{1}n_{O} &\rightarrow \text{fission fragments} \\ &+ 2 \text{ to } 3 \text{ neutrons} + \text{energy} \end{aligned} \tag{5}$$

In a nuclear reactor, the heat generated by this process is used to produce steam for powering a turbine generator. Conventional reactors typically use fuel rods that are enriched with ^{235}U to a level of 2–4%. A portion of the neutrons produced in fission are sorbed by the more plentiful ^{238}U, eventually leading to the formation of fissile ^{239}Pu, contributing about one-third of the energy produced in the reactor. Within about 3 years the buildup of fission products within the nuclear fuel interferes with the transfer of neutrons, reducing the efficiency to the point where the fuel must be removed [10, 13]. There are two groups of fission fragments: relatively short-lived, highly radioactive isotopes (^{137}Cs and ^{90}Sr) and longer-lived isotopes (^{99}Tc, ^{135}Cs, and ^{129}I). Both groups are major contributors to radioactive waste streams, and need to be assessed in terms of their behavior in the environment when developing a waste disposition policy.

Radionuclide Sources

Natural Sources

Some natural radionuclides are primordial in origin (e.g., ≈ 10 ppm Th and ≈ 4 ppm U) with half-lives generally exceeding 10^{9} years. Other natural radionuclides are generated by continuous bombardment of the earth and atmosphere by cosmic rays, including ^{14}C which is important in radiometric dating, as well as ^{3}H, ^{7}Be, and ^{10}Be (see Table 1) [10, 12]. In addition to naturally occurring radionuclides, anthropogenic sources in terrestrial systems include fallout from the atmospheric testing of nuclear weapons, nuclear material mining, and processing activities for defense and industrial uses, nuclear power generation and fuel reprocessing and disposal, and biomedical wastes. Some recognition of natural background radiation levels is essential in defining the boundary of contaminated sites and establishing an appropriate threshold for cleanup levels [16].

Uranium mining and milling efforts generate the largest volume of radioactive waste, often containing

Radionuclide Fate and Transport in Terrestrial Environments. Table 1 Source, characteristics and examples of the various nuclide classes [10, 12]

Class	Sources	Characteristics	Examples
Stable	Found in nature		^{12}C, ^{14}N, ^{16}O
Primary natural radionuclides	Initial forming of the universe	≈26 known radionuclides with generally very long half-lives	^{40}K (1.28 × 10^9 y), ^{238}U (4.47 × 10^9 y), ^{235}U (7.13 × 10^8 y)
Secondary natural radionuclides	Decay products of natural radionuclides	≈38 known; short half-lives	^{226}Ra (1,600 y), ^{234}Th (24.1 d)
Induced natural radionuclides	Cosmogenic formation in troposphere	≈10 known	^{3}H (12.3 y), ^{14}C (5730 y), ^{7}Be, ^{10}Be, ^{22}Na, ^{32}P, ^{33}P, ^{35}S, ^{39}Cl
Anthropogenic radionuclides	Man made	≈2,000 known	^{60}Co (5.3 y), ^{137}Cs (30.2 y)

Values in parentheses represent radioactive half-lives.

additional hazardous metals and sulfides, which generate acidity upon oxidation that enhances contaminant release. However, the relative environmental footprint associated with U mining for nuclear energy is orders of magnitude less than coal for an equivalent unit of energy production. Under the Uranium Mill Tailings Remedial Action (UMTRA) program, the US Department of Energy (DOE) has made considerable progress in the remediation of 24 U mining and milling sites in 10 states. Nonetheless, U mining and milling will likely increase within the next decade as current fuel stockpiles are exhausted [4, 17, 18]. Other important sources of naturally occurring radioactive materials (NORM) include waste generated from the phosphate fertilizer industry, the scales from oil and gas extraction, ash from conventional coal-based power generation, and the waste from metal mining and processing facilities [19].

Nuclear Weapons Development, Testing, and Production

Since the first nuclear test in New Mexico in 1945, more that 2,400 nuclear test explosions have occurred, including more than 500 atmospheric tests, resulting in the global dispersal of ^{239}Pu, ^{137}Cs, ^{14}C, ^{90}Sr, ^{131}I, ^{3}H, and much lower quantities of numerous short-lived radionuclides. The Former Soviet Union (FSU) conducted more than 460 nuclear tests at the Semipalatinsk test site in Kazakhstan [7, 18, 20]. The total radioactivity released to the atmosphere from nuclear weapons testing has been estimated to be over 2 × 10^8 TBq [21]. Even now, there are still approximately 27,000 nuclear weapons spread among the various nuclear powers, and the USA and Russia each have an inventory of about 40 metric tons of Pu [22].

Beyond the releases associated with weapons testing, nuclear weapons development and production has resulted in the generation of large stockpiles of nuclear waste and extensive contaminated areas [23], the disposition of which is still in question. For example, about 2 million m^3 of high-level waste currently stored in 177 underground tanks was produced at the two Pu processing lines at the DOE Hanford Facility, with 67 of the tanks having leaked at some point [7]. In another example, considerable environmental contamination occurred at the Mayak nuclear facility in the FSU, with substantial volumes of accumulated high-level waste (>350 million Ci) awaiting final disposition [24].

Nuclear Energy

Currently, about 17% of the world's electricity is derived from nuclear power. A 1,000 MWe nuclear reactor generates about 27 metric tons of waste a year (globally ≈ 12,000 metric tons year^{-1}), compared to the 400,000 metric tons of waste generated by a coal-fired plant producing the same amount of electricity [17]. The USA currently has 103 of the 443 nuclear power reactors worldwide, France has 59, and Japan has 54, with much of the expansion in nuclear power generation set in Asia and the developing world.

The three major producers combined represent about 57% of the world's total electrical production from nuclear power. However, France (70% of domestic electricity), Slovakia (57%), and Belgium (54%) are the countries most dependent on nuclear power [7]. Most if not all of the nuclear power reactors in the USA have or will petition the NRC to extend their initial 40-year operational limit.

More than 200 radionuclides are produced during the operation of a typical nuclear power reactor. As discussed previously, reactor fuel must be routinely replaced because of fission product buildup that disrupts neutron transfer, inhibiting fission and energy production. Many of the SNF radionuclides have relatively short half-lives and decay within the first few decades [7]. Important fission products associated with the SNF include ^{135}Cs, ^{137}Cs, ^{129}I, ^{90}Sr, and ^{99}Tc, representing about 3% of the residual SNF weight. The fission products ^{137}Cs and ^{90}Sr account for much of the SNF radioactivity for the first 10–100 years of storage. Despite the public's reservations concerning the safety of nuclear power, controlled and uncontrolled releases of radioactivity from the operation of commercial nuclear reactors represents less than 3% of the amount released from atmospheric nuclear weapons testing [21]. Furthermore, the atmospheric emission of radioactivity in the form of U and Th from a conventional coal-fired plant can exceed the atmospheric release from nuclear power when compared on an equivalent energy output basis [25, 26].

In Europe and Japan, much of the SNF fuel is reprocessed to remove the fissile material (U and Pu) for subsequent reuse. For instance, the French facility at La Hague recycles SNF for France, Germany, Belgium, Switzerland, and the Netherlands, with the recovered waste returned to the originating country for disposal. Another commercial SNF reprocessing facility serving multiple countries is housed at the Sellafield nuclear facility in England, with additional reprocessing facilities in India, Japan, Russia, and China [7]. However, fuel reprocessing is not without its limitations and drawbacks. Contaminants such as Cm and Am build up in the recycled SNF and limit continued reprocessing. Eventually, both high-level waste removed during reprocessing and the recycled SNF will require disposal in a geologic repository [22]. Although deep geologic containment has largely been accepted as the primary disposal strategy for commercially derived high-level waste, no country to date has completed the design and construction of such a repository [7]. Given the limited mass of ^{239}Pu required to create a nuclear bomb ($\approx$7–10 kg), the potential diversion of fissile materials in SNF destined for reprocessing is viewed as a potential route to expedited nuclear weapons proliferation.

In contrast to Europe, the USA currently practices a once-through nuclear fuel cycle due to concerns over proliferation, with much of the SNF remaining in storage pools housed on the reactor sites awaiting the opening of the Yucca Mountain repository or the choice of an alternate disposal strategy. However, the USA is constructing a facility for combining the surplus inventory of Pu originally produced for nuclear weapons with U to produce a mixed-oxide fuel for use in commercial nuclear reactors. In addition, fourth-generation nuclear reactors currently in development offer several potential advantages, including a closed fuel cycle that produces little or no long-lived radioactive waste and reduces the generation of materials that aid in nuclear proliferation, i.e., enriched U and ^{239}Pu [27].

Nuclear Accidents

Accidental discharges at both government and commercial nuclear facilities have played an important role in shaping public opinion concerning the relative safety of nuclear energy. In 1957, a fire within the graphite core of a reactor at the British nuclear facility at Windscale, known now as Sellafield, released significant amounts of ^{131}I and ^{137}Cs to the countryside. The Windscale facility produced ^{239}Pu for the British nuclear weapons program. Following the fire, a second Windscale reactor of similar design was shut down, and the damaged reactor unit was essentially decommissioned in place. The same year, the Kyshtym disaster, known for the name of a local town, occurred at the Mayak nuclear facility, the center for nuclear weapons production in the FSU. The cooling system associated with a series of nuclear waste storage tanks failed, resulting in a nonnuclear explosion that released significant levels of ^{90}Sr and ^{137}Cs throughout the adjacent countryside known since as the East-Ural Radioactive Trace [24, 28].

Three Mile Island and Chernobyl

In 1979, a mechanical failure in the cooling water loop at the Three Mile Island (Dauphin County, PA) reactor resulted in a lowering of the water level within the reactor core which exposed the upper section of the fuel rods. While nuclear fission within the reactor ceased in the absence of the water moderator, the temperature of the power rods increased rapidly due to the ongoing decay of existing fission products, resulting in a partial meltdown of the fuel rods. Contaminated water from the primary cooling loop was released within the plant facility, and noble gases (e.g., ^{135}Xe and Kr) and fission products (^{129}I and ^{131}I) were released into the environment. There was concern at the time that the zirconium alloy coating the rods would react with the steam to produce H_2 that could potentially react with O_2 to produce an explosion, resulting in additional dispersion of reactor fuel materials and fission products in the environment [10].

Another serious accident occurred in 1986 at the Chernobyl power station near Kiev, Ukraine, when the control rods were removed from the reactor during a systems test, accelerating the chain reaction. The buildup of heat eventually vaporized the cooling water, causing the core to explode. As in the accident at Three Mile Island, nuclear fission ceased due to the lack of the moderator, but the graphite within the exposed reactor ignited, releasing additional transuranics (≈ 800 Ci ^{238}Pu, ≈700 Ci ^{239}Pu, ≈1,000 Ci ^{240}Pu) and fission products into the environment, i.e., ^{137}Cs, ^{134}Cs, ^{90}Sr, and ^{131}I. Because of the danger in approaching the burning facility, the fire was eventually extinguished using helicopters and the reactor was entombed in concrete [10, 29].

Soils as the Repository for Radioactive Waste

Reactive Soil Components

Excluding radioactive decay, the major chemical, physical, and biological processes governing the fate and transport of radionuclides in the soil system include aqueous speciation/complexation, sorption/desorption, convection/dispersion, resuspension and erosion, and redox state and other chemical species transformations. In terms of biological reactions, radionuclides may be assimilated/immobilized into the living organic matter within the soil microbiota and absorbed by the roots of higher order plants as the first steps in trophic transfer in terrestrial systems.

Before discussing the reactions controlling the fate and transport of specific radionuclides of interest, it is important to review the physical and chemical properties of soils that control reactivity. Soils consist of three phases; the solids account for about 40–50% of the soil volume, and the aqueous and gaseous phases account for the remaining balance. As the degree of saturation increases, the relative volume of the gaseous phase decreases. The degree of saturation is an important variable controlling solute fate and transport in soil and vadose systems, including advective/dispersive transport and the buildup and transfer of volatile constituents that impact biotic activity and soil redox state (i.e., O_2, CO_2, NH_3, CH_4, etc.). Although the term “soil” is generally restricted to the top meter or so of the earth’s crust, the current discussion concerning various soil processes controlling radionuclide fate and transport is generally relevant to the vadose zone and saturated groundwater systems.

Phyllosilicate Clays

Surface area is a major factor controlling the reactivity of a solid phase, including soil materials. Therefore, discussion will focus on the clay fraction (<2 μm) that tends to dominate soil reactivity because of the large surface to volume ratio. However, the term “clay” can also be used in reference to a specific soil textural class or classes based on the relative proportion of sand (2 mm −50 μm), silt (<50 μm to 2 μm), and clay-sized (<2 μm) materials, or a specific type of mineral that consists of repeating silicate sheets, i.e., phyllosilicate clays. All three definitions are relevant to the current discussion of soil reactivity.

Minerals are generally classified based on the dominant anion or anionic group within the structure, e.g., halides, carbonates, sulfides, phosphates, oxides/hydroxides/oxyhydroxides, and silicates [30]. Much of the current discussion will focus on phyllosilicates and metal oxides. However, other soil mineral classes will be discussed in relation to processes controlling the behavior of specific radionuclides as warranted. Readers are directed to Dixon and Schulze [31] for a more thorough introduction to the complexities of soil mineralogy.

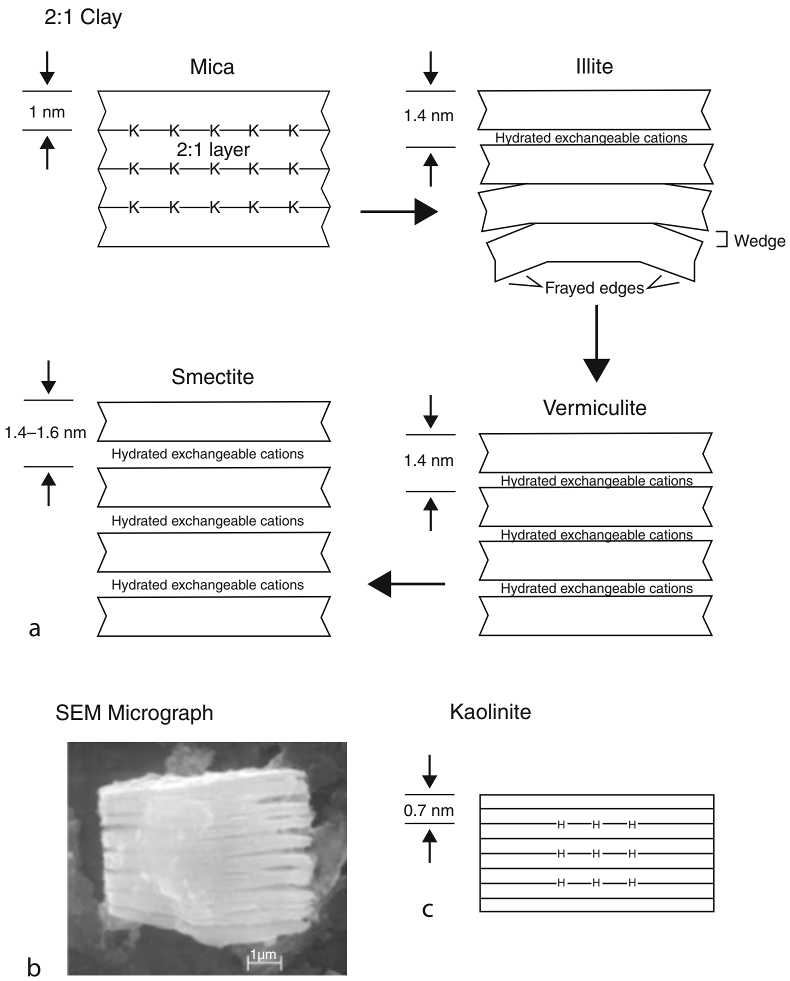
Silicates are an extremely diverse mineral class accounting for well over 90% of the earth's crust. In discussing clay structure, the octahedral and tetrahedral layers will be referred to as sheets, while the term "layer" refers to the fundamental clay unit representing the combination of the sheets. The fundamental structural unit for silicates is the SiO_4 tetrahedron, consisting of four O^{2-} ions surrounding a single Si^{4+} ion. For the layer silicates, the three basal oxygens are shared with other Si-tetrahedral units to form a sheet. The second primary structural unit in phyllosilicate clays is the octahedron consisting of Al^{3+} (also Fe^{2+} and Mg^{2+}) surrounded by six hydroxyl groups, OH^-, which also combine to form a sheet structure [31].

Phyllosilicate clays are divided into 2:1 and 1:1 type minerals based on the number and sequence of tetrahedral and octahedral sheets in the layer structure. The 2:1 clay minerals consist of an octahedral sheet sandwiched between and sharing the apical O^{2-} ions of two Si-tetrahedral sheets. The 2:1 clay minerals display a net surface charge due to imperfections in the crystal structure derived from the substitution of a lower valence cation for a higher valence cation, termed isomorphic substitution. Isomorphic substitution generally results from the substitution of Al^{3+} for Si^{4+} in the tetrahedral sheet and Mg^{2+} or Fe^{2+} for Al^{3+} in the octahedral sheet, resulting in diffuse permanent negative charge that is compensated by the attraction of common alkali and alkali-earth cations to the clay surface, i.e., cation exchange capacity (CEC). Isomorphic substitution is insensitive to conditions in the solution environment, and therefore termed constant charge. In addition to the external siloxane oxygens of the basal tetrahedral planes for each layer, terminal aluminol (AlOH) and silanol (SiOH) groups are found along the edge boundaries of the clay layers. In addition to the diffuse cation exchange capacity resulting from isomorphic substitution, the edges can serve as sorption sites depending on pH and other solution conditions, just like other variable-charge sites to be discussed below.

The 2:1 mineral class can be further categorized based on the location and degree of isomorphic substitution within the mineral structure, i.e., micas (illites), vermiculites, and smectites (Fig. 1a). The terms "illite" and "hydrous mica" are often used to refer to weathered micas. Layer charge is defined as the magnitude of charge derived from isomorphic substitution per mineral formula unit for a given clay, with the layer charge for mica > illite > vermiculite > smectite. However, the general distinction between these groups of 2:1 clays and the 1:1 kaolin group discussed below is operationally defined by their response in terms of interlayered spacing to various cation saturation, heating, and desiccation treatments, as determined by x-ray diffraction (XRD) according to standardized criteria [32].

Increasing layer charge results in increasing interlayer bonding as adjacent clay layers tightly share interlayer cations that balance the charge resulting from isomorphic substitution. For instance, smectites have a relatively low layer charge when compared to vermiculites and micas, with much of the charge residing in the octahedral layer; however, smectites display a higher CEC and reactive surface area than mica because cations, water, and other polar molecules can be exchanged between the clay interlayer and the bulk solution due to the weaker interaction of adjacent clay layers (Fig. 1a). In contrast, micas have a high degree of isomorphic substitution within the tetrahedral layer, resulting in more localized charge distribution, but display limited CEC. Cations with ionic radii similar to the ditrigonal cavity formed by six corner sharing Si tetrahedra present on the exterior planar surface of 2:1 clay minerals (i.e., K^+, $NH_4{}^+$, and Cs^+) are strongly bound between mica layers because of the close proximity to the origin of charge in the tetrahedral layer. This strong interaction, represented by K in Fig. 1a, limits the ability of water or cations from the bulk solution to penetrate the mica clay interlayer. Vermiculites display an intermediate behavior, with cations like K^+, $NH_4{}^+$, and Cs^+ held within the interlayer becoming non-exchangeable under certain conditions.

Although larger than a clay-sized particle (i.e., $> 2\ \mu m$), the clay mineral displayed in the SEM micrograph Fig. 1b can serve to illustrate the heterogeneous surfaces present on clay minerals, despite the fact that sorption processes occur at a much more refined scale than apparent in this macroscopic representation of various clay features. Distinguishing various external planar, interlayer, and edge sites will become important when discussing ^{137}Cs and ^{90}Sr behavior in soil.



Radionuclide Fate and Transport in Terrestrial Environments. Figure 1
(**a**) Weathering transition and interlayer expansion properties of 2:1 clays from mica to smectite. (**b**) SEM micrograph of 2:1 clay illustrating planar surface, interlayer and edge character apparent at a larger scale (20 kV, Seaman unpublished). (**c**) Kaolinite clay with no interlayer expansion because of hydrogen bonding between 1:1 sheets (A and C revised from NCRP [156])

Due to their high reactivity and surface area, and swelling properties, smectites have been evaluated for use as engineered barriers or back-fill materials for proposed high-level nuclear waste repositories (HLNWR) in Japan and Europe. However, recent evidence suggests that high levels of radiation (i.e., alpha recoil and ionizing radiation from fission products) can induce amorphization, altering both the solubility and sorptive properties of the smectites [33]. Therefore, the impact of radiation on natural environmental sorbents, such as soil clays, warrants greater research attention.

In some highly weathered soils like those of the southeastern USA, polymeric $Al\,(OH)_n^{(3-n)+}$ sheets often occupy portions of the interlayer of 2:1 clays, blocking cation exchange sites and limiting interlayer expansion and collapse in response to cation saturation and hydration conditions. These clays are commonly classified as hydroxyl-interlayered smectites (HIS) or hydroxyl-interlayered vermiculites (HIV) based on the degree to which interlayer expansion and collapse are restricted [34]. Although the $Al(OH)_n^{(3-n)+}$ polymers reduce CEC and limit interlayer collapse, their presence can result in sorption sites that potentially display unique preferences for soil contaminants.

In 1:1 phyllosilicates there are three anion planes: the basal siloxane plane associated with the tetrahedral sheet, the middle layer associated with O^{2-} ions within both the tetrahedral and octahedral sheets, including apical tetrahedral O^{2-} ions shared with the octahedral sheet, and the hydroxyls associated with the remaining planar surface of the octahedral sheet (Fig. 1c). Kaolinite consists of a series of stacked 1:1 structural layers with Al^{3+} in the octahedral sheets and Si^{4+} in the tetrahedral sheets. Although structurally neutral due to limited isomorphic substitution, the adjacent 1:1 layers are restricted from expanding by hydrogen bonding associated with the aluminol planar surface of the octahedral sheet and the basal siloxane surface of the next kaolinite layer. In general, surface charge in kaolinite is derived from the amphoteric functional groups on the edges of the clay layers.

Zeolites

Zeolites are framework aluminosilicates consisting of extended three-dimensional networks of linked SiO_4 and AlO_4 tetrahedra. Zeolites derive cation exchange capacity from Al^{3+} substitution for Si^{4+} with the size of the channel determining the type of cation that is preferred [35, 36]. In nature, zeolites are found in significant concentrations associated with sedimentary materials of volcanic origin [37]. Although not common in soils, zeolites are important because of their use as selective sorbents for removing radionuclides from waste streams, treating contaminated systems, and backfilling to buffer radioactive waste storage systems [38]. Zeolites possess unique interconnected channels or voids that form selective sorption sites for water and alkali and alkali-earth metals, including $^{135/137}Cs^+$ and $^{90}Sr^{2+}$ [39]; however, zeolites display a limited sorption affinity for metals that tend to hydrolyze or form strong complexes with carbonate (e.g., Np, Th, and U) [39–41].

Metal Oxides and Hydroxides

Soil minerals are further characterized as either being primary or secondary in nature, with primary referring to minerals that were formed at elevated temperatures, and then derived from igneous or metamorphic rock. Secondary minerals are formed at lower temperatures and derived from sedimentary minerals or soil weathering processes. As primary soil minerals break down and weather, ionic constituents are released into solution that may recombine and form other minerals, i.e., secondary minerals. Iron, Al, Mn, and Ti released during the weathering of primary minerals tend to persist and form stable oxides/oxyhydroxides that are resistant to continued weathering, with the metal cation occupying octahedral sites [42, 43]. In general, the term "oxide" will be used in reference to any metal-oxide, hydroxide, or oxyhydroxide mineral.

Gibbsite (γ-$Al(OH)_3$) is the most common Al hydroxide found in soils. Gibbsite is structurally analogous to the $Al(OH)_n^{(3-n)+}$found in the interlayers of HIS and HIV, and consists of parallel layers of Al in octahedral coordination with OH anions, with every third cation center remaining vacant. Ion adsorption by gibbsite occurs at the edge surfaces of the layer sheets. Goethite (α-FeOOH) and hematite (α-Fe_2O_3) are the most common Fe oxides found in soils and sediments. The poorly ordered ferrihydrite ($Fe(OH)_3$) is also quite common, but its limited crystallinity precludes identification based on XRD. However, the crystalline and amorphous Fe oxide contents for soils are typically estimated by selective extraction using the citrate-dithionite-bicarbonate (CDB) and ammonium-oxalate (AO) extractions, respectively [44]. Several metal cations, including Al^{3+}, have the ability to substitute for Fe within Fe oxide structures [45].

Although generally found in lower concentrations than Fe and Al oxides, Mn(III/IV) oxides are strong sorbents and highly important in surface-mediated redox reactions [46, 47]. Manganese oxides are difficult to study because of their complicated mineralogy and highly variable structure, with various stable and metastable phases found in soils.

Organic Matter

Soil organic matter (SOM) is a complex, polymeric mixture of aliphatic and aromatic compounds derived from the decomposition of plants and animals by heterotrophic microorganisms. SOM possesses a wide range of organic functional groups that bind with radionuclides, including hydroxyl (OH), carboxyl (COOH), sulfite (SO_3), sulfide (S), and amine (NH_2) functionalities. Soil organics can enhance the dissolution of minerals through the complexation of metals, and can prevent the formation of Fe and Al oxides by forming strong organic complexes that hinder precipitation. Binding to SOM also plays a large role in controlling the bioavailability and toxicity of metals [48]. Furthermore, SOM can facilitate redox transformations at mineral surfaces, and alter the charge properties of soil clays, generally enhancing soil aggregation [49].

SOM can occur in the particulate phase, colloidal phase, or dissolved in solution. This division is operationally defined by the methods used to separate the fractions. Typically, 0.45 μm is the filter pore size used to separate particulates from what is traditionally considered "dissolved." The filtrate can be further separated to isolate the colloidal fraction. The colloidal fraction can play a significant role in the mobilization/immobilization of soil contaminants [50–52]. For example, U-humic colloids have been observed to constitute $>$ 75% of the soluble U found in groundwater around the Gorleben salt domes in Germany [53, 54]. Microbes and microbially produced substances also play a significant role in the migration of radionuclides because of their influence on redox status which can affect speciation and thus the ultimate fate of redox-sensitive radionuclides, such as U and Pu, in the environment [55].

Important Soil Processes

Radionuclide Partitioning

In discussing soil processes, the term "sorption" is used to describe a decrease in the concentration of a given constituent in the aqueous phase, in this case a specific radionuclide, without implying a specific partitioning mechanism, such as surface complexation, precipitation, or ion exchange. The use of such a generic term is a reflection of the complexity of the soil environment, and the difficulty in experimentally distinguishing between an array of "sorption" mechanisms using standard wet-chemical analysis techniques [56–58]. In recent years, Monitored Natural Attenuation (MNA) has received considerable attention as an alternative to more invasive remediation strategies that can be highly destructive to functioning, contaminated ecosystems, even for addressing radionuclide contamination [59]. However, a thorough understanding of radionuclide partitioning mechanisms is critical to the effective application of MNA for contaminated sites where expensive, invasive remediation is unlikely to greatly lessen environmental risks.

Chemical Speciation

Radionuclides released in global fallout display a wide range of chemical properties that impact their behavior in the environment, including both cationic and anionic species and redox-sensitive elements. Initially, radionuclide speciation in the environment is highly dependent on the form in which it was released, often becoming less labile over time. Some of the actinides display very complex redox speciation (e.g., Pu), with the possibility for multiple redox states to coexist. Further, fission products resulting from nuclear power generation represent every group in the periodic table [10]. The form of the radionuclide (i.e., chemical species) and the physical and chemical properties of the soil materials largely control the fate and transport of the contaminant, including solid-phase partitioning, bioavailability, and subsequent trophic transfer [58, 60–62]. When limited data are available, the understanding of radionuclide behavior in the environment is based on that of a biogeochemical analogue/surrogate, either an element with similar chemical properties (i.e., ^{137}Cs vs K) or a stable or longer-lived isotope of the element of concern, like ^{135}Cs for ^{137}Cs [61, 63, 64].

Most common analytical techniques, such as mass- or radiolytic-based methods, provide estimates of total isotope concentration, but are insensitive to chemical speciation. Spectroscopic methods, such as x-ray absorption spectroscopy, are generally required to determine chemical speciation [65]. Goldberg et al. [58] provide an extensive review of spectroscopic techniques for evaluating structural features of both sorbent surfaces and adsorbed species. Such information is critical in validating a mechanistic description of reactive solute partitioning. However, the detection limits for solid-phase speciation techniques are often inadequate for their application to radionuclide concentrations associated with environmental samples, and special care must be taken to ensure that various changes in chemical speciation do not occur as a result of sample collection and storage before analysis, or even as a result of the analysis method itself. Colloidal suspensions are particularly sensitive to artifacts associated with storage and handling [66, 67].

Sequential extraction procedures, such as the widely used Tessier et al. [68] and Miller et al. [69] methods, can provide information concerning contaminant partitioning and bioavailability. Such methods were first developed for the solid-phase speciation of metals, using a series of progressively more-aggressive extractants to isolate metals or radionuclides presumably associated with various contaminant partitioning mechanisms, i.e., water soluble, ion exchangeable, sorbed to Fe oxides, etc.; however, sequential extraction methods are operational in nature and suffer from limitations associated with the non-specificity of extractants and problems related to re-adsorption of metals during the extraction process [70, 71]. From an environmental assessment standpoint, sequential extractions can provide useful information concerning specific conditions under which a contaminant may be mobilized (i.e., acidic, reducing, etc.), and evaluating changes in solid-phase associations resulting from a given remediation treatment [72–74].

Cation Exchange Reactions

Cation exchange is one of the most common types of sorption reactions occurring in soils [75]. Using Sr^{2+} as an example, a typical cation exchange reaction can be formulated as:

$$2(R^{-} - Na^{+}) + Sr^{2+}_{aq} \Leftrightarrow (R^{2-}_{2} - Sr^{2+}) + 2Na^{+}_{aq} \qquad (6)$$

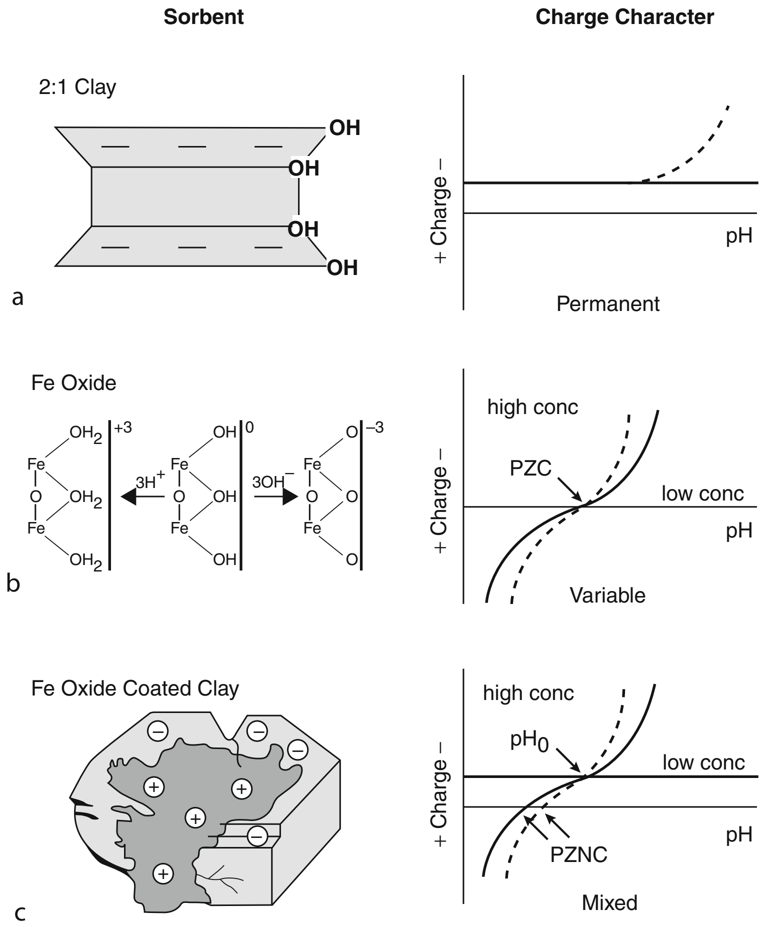
where R^- represents an exchange site associated with the solid phase, and Na^+ and Sr^{2+} are exchangeable cations. Such reactions are quite rapid and fully reversible. Although cation exchange can be written as mass action equations, the resulting selectivity coefficients are not thermodynamic constants because of the difficulty in correcting for activities associated with the exchanger phase [76]. The CEC of a soil material is generally assumed to represent the charge associated with isomorphic substitution within phyllosilicate clays plus the exchange capacity associated with SOM. CEC is generally reported in units of equivalent cations sorbed per unit mass of soil, i.e., centi-moles of charge per kilogram of soil ($\mathrm{cmol}_{(+)}\ \mathrm{kg}^{-1}$).

For cation exchange sites resulting from isomorphic substitution, the relative exchange affinity increases with valence, i.e., bivalent cations are sorbed preferentially to monovalent cations. However, the selectivity tends to decrease with increasing ionic strength. Thus, dilution increases the bivalent to monovalent cation ratio on the exchange complex. For cations of the same valence, the relative affinity for cation exchange sites increases with the non-hydrated radii of the ion: $Cs^{+} > K^{+} > Na^{+} > Li^{+}$ and $Ba^{2+} > Sr^{2+} > Ca^{2+} > Mg^{2+}$. However, differences in cation selectivity as a function of ionic size indicate that additional factors beyond coulombic attraction contribute to such preferences.

In Fig. 2a, a simplified diagram of a 2:1 clay is presented with the negative signs representing the charge (i.e., cation exchange capacity) derived from isomorphic substitution. The right side of Fig. 2a is a diagram illustrating the impact of pH on surface charge (solid line). Since the charge is derived internally, it remains unchanged with variations of soil pH. However, terminal OH functional groups located on the edges of 2:1 clays are responsive to soil pH, and an increase in net negative charge (i.e., CEC) can be observed with increasing pH. The phyllosilicate edges can also act as "specific" sorption sites for both cationic and anionic species as discussed below for other variable-charge surfaces.

Variable Charge and Surface Complexation

In contrast to charge resulting from isomorphic substitution, amphoteric surfaces on soil minerals and ionizable functionalities on SOM are sensitive to the conditions in the solution environment (i.e., pH, ionic strength, counterion valence, etc.), and therefore termed variable-charge sites/minerals. The oxides and oxyhydroxides of Al, Fe, Ti, Mn, and other metals display this type of charging mechanism, e.g., $\equiv M - OH$, where M represents the primary metal and OH is a surface hydroxyl filling the coordination sphere of the metal (Fig. 2b). The surface charge of the oxide becomes more positive with increasing acidity, and more negative with increasing alkalinity. As noted before, the silanol and aluminol edges of phyllosilicate clays (Fig. 2a) also display this type of charge behavior, with the CEC for kaolinite largely derived from amphoteric edge sites because of the limited isomorphic substitution within the clay structure.



Radionuclide Fate and Transport in Terrestrial Environments. Figure 2
Common mineral sorbent types found in soils and electrostatic charge expression as a function of solution conditions (i.e., pH and ionic strength): (**a**) constant charge, (**b**) variable charge, and (**c**) a combination of constant and variable-charge types

For variable-charge minerals, the reactivity of the mineral surface will depend on the structure of the metal hydroxide, the mineral surface area and surface reactive site density, and the point of zero charge (PZC). The PZC (pH_{pzc}) is defined as the pH at which positive and negative surface charges are equal, i.e.,

$$\equiv M - OH_2^+ \quad \text{equals} \quad \equiv M - O^- \qquad (7)$$

which also corresponds to the pH value for the minimum solubility of the metal-oxide/oxyhydroxide in question, and to conditions that facilitate the rapid coagulation of oxide suspensions [76]. Conceptually, the acid/base equilibrium at the mineral surface can be represented by the following:

$$\equiv M - OH_2^+ \Leftrightarrow \equiv M - OH^0 + H^+ \qquad K_{a1} \qquad (8)$$

$$\equiv M - OH^0 \Leftrightarrow \equiv M - O^- + H^+ \qquad K_{a2} \qquad (9)$$

where K_{a1} and K_{a2} are the apparent acidity constants, with the surface charge becoming more positive as the pH decreases and more negative as the pH increases (Fig. 2b). In contrast to an amphoteric solute, the mineral surface provides a degree of structural constraint to the acid/base functionalities that alters the apparent dissociation constants, constraining subsequent protonation and deprotonation, causing the dissociation constants to express more variable amphoteric character. At a specific pH, however, the surface charge of a variable-charge mineral will increase with increasing ionic strength, as illustrated by the two charge curves on Fig. 2b.

As a consequence of this amphoteric behavior, both anions and cations may be electrostatically sorbed to the charged oxide surface. Such electrostatic bonds in which the sorbed ion remains hydrated are termed "outer sphere" complexes. Using the same convention, hydrous oxides can form stronger "inner sphere" surface complexes with metal ions resulting in the displacement of a proton and the generation of excess positive charge, which may in term attract an anion:

$$\equiv Fe-OH+Pb^{+2}\Leftrightarrow\equiv Fe-OPb^{+}+H^{+} \tag{10}$$

A similar reaction can be written for the "inner sphere" sorption of an anion by ligand exchange with a surface OH^- or H_2O group:

$$\equiv Fe-OH+A^{n-}\Leftrightarrow\equiv Fe-A^{(n-1)}+OH^{-} \tag{11}$$

where A^{n-} represents the anion in solution. As indicated by the previous equations, surface complexation reactions are competitive processes that depend on the pH and the sorbate in question. Such reactions will be discussed further in the context of Surface Complexation Modeling (SCM).

It is important to note that soils generally contain a combination of constant charge clays, variable-charge oxides, and SOM with various reactive functionalities. Figure 2c represents a phyllosilicate clay impacted by an oxide coating. Although the term "coating" is widely used to describe the interactions between Fe oxides and phyllosilicate clays, it is somewhat misleading as the Fe oxides often display a very discrete, fine-grained morphology that is best described as armoring the clay surface [77]. As the charge diagram on the right illustrates (Fig. 2c), a fixed level of negative charge derived from isomorphic substitution is present at all times, with an additional variable-charge component that is responsive to pH and ionic strength. In such systems, the variable-charge oxides may actually block a portion of the CEC in a manner similar to the way $Al(OH)_n^{(3-n)+}$polymers can block interlayer exchange sites in 2:1 clays. For systems with limited CEC, the pH conditions may be such that the positive charge (anion exchange capacity, AEC) derived from the oxides may be equivalent to or even exceed the CEC.

Redox Reactions

By convention, redox reactions are written in terms of the reductive half reaction. For instance, the reduction of Fe(III),

$$Fe^{3+}+e^{-}\Leftrightarrow Fe^{2+} \tag{12}$$

where e^- represents the electron. To be complete, reactions such as Eq. 12 must be balanced by an appropriate oxidation reaction:

$$4H^{+}+O_{2(g)}+4e^{-}\Leftrightarrow 2H_2O \tag{13}$$

Reaction 13 is reversed and balanced to represent the overall redox reaction:

$$Fe^{3+}+\frac{1}{2}H_2O\Leftrightarrow Fe^{2+}+H^{+}+\frac{1}{4}O_{2(g)} \tag{14}$$

with the electrons from the two half reactions canceling each other out [13]. In a more relevant example, the two prominent redox states of U (IV and VI) in soils are linked by

$$UO_2^{2+}+4H^{+}+2e^{-}\Leftrightarrow U^{4+}+2H_2O \tag{15}$$

with the subsequent precipitation of the sparingly soluble $UO_{2(S)}$ [78]. Uranium (VI), originally thought to be reduced abiotically, can also be reduced by microbially mediated processes [79, 80]. For example, microbially produced acetate can reduce U(VI) to U(IV) as seen in reaction 16.

$$\begin{aligned}&CH_3COO^{-}+4U(VI)+4H_2O\rightarrow 4U(IV)\\&\quad+2HCO_3^{-}+9H^{+}\end{aligned} \tag{16}$$

Redox conditions may control the relative solubility of a contaminant by altering its chemical speciation and/or the properties of the sorptive phase. For instance, the reduced form of U, U(IV), is much less soluble than the oxidized form, U(VI) [13]. Furthermore, reducing conditions may lead to the dissolution of important sorbents such as goethite (α-FeOOH), releasing both adsorbed contaminants as well as contaminants that may have substituted for Fe within the mineral structure.

Precipitation/Dissolution

Dissolution and precipitation reactions are likely to dominate in highly contaminated systems, such as waste disposal facilities, or where steep pH and oxidation/reduction gradients persist [81, 82]. A full treatment of mineral equilibrium interactions is beyond the scope of the current chapter. For more detail the reader is directed to references [76] and [83]. The Gibbs free energy of a reaction is

$$\Delta G_r = RT \ln \frac{Q}{K_{eq}} \tag{17}$$

where R is the gas constant, T is the standard temperature (K), Q is the ion activity product (IAP) of soil solution components, and K_{eq} is the equilibrium product, i.e., the solubility product (K_{sp}) for the precipitate of interest. At equilibrium for a given system, $Q = K_{\mathrm{sp}}$. The saturation index (SI) with respect to a specific solid phase is defined as

$$SI = \log_{10}\left(\frac{Q}{K_{eq}}\right) \tag{18}$$

A SI $<$ 0 (i.e. negative value) indicates that the solution is under-saturated, while a SI $>$ 0 indicates the solution is supersaturated with respect to a specified solid phase. A SI of 0 indicates that the solution is in equilibrium with a given solid.

At equilibrium with a solid phase, the aqueous-phase activity of the soluble components is independent of the amount of the precipitate present. In contrast to adsorption reactions, precipitation involves nucleation and then solid-phase growth in three dimensions. Coprecipitation occurs when an element is incorporated within a precipitate as a trace or minor component, with metal substitution readily observed for Fe oxides [45]. In a relevant example, the precipitation of U(VI) with secondary Al-phosphates has been observed when apatite ($Ca_5(PO_4)_3OH$) was added to an ultisol contaminated with U and other metals, rather than sorbing to residual apatite or the formation of a discrete U phosphate [74, 84].

Before concluding the detailed discussion of soil mineralogy and reactive sites, it is important to note that the properties of natural soil minerals are often quite different from their pure mineral specimens or synthetic analogues on which much of the fundamental contaminant partitioning studies are based. Recent studies have elucidated the role of less abundant mineral and organic substrates as important surface chemical modifiers and have demonstrated the complex coupling of reactivity between permanent-charge phyllosilicates and variable-charge Fe-oxyhydroxide phases [77, 85–87].

Models for Describing Radionuclide Sorption and Plant Uptake

Empirical Partitioning Isotherms

Partitioning isotherms are often used to describe solute behavior in recognition of the simplistic, empirical methods by which solute–surface interactions are commonly evaluated. The three most commonly used equilibrium sorption models include the Linear, Freundlich, and Langmuir equations. When used to describe solute transport, the sorption behavior described by the three isotherms is assumed to be both rapid and fully reversible. The Linear isotherm equation is

$$S = K_d C \tag{19}$$

where S is the amount sorbed, K_d is the linear distribution/partition coefficient (L kg^{-1}), and C is the solute concentration in equilibrium with the sorbent phase. For radionuclides, the K_d is often calculated in terms of Bq within both the aqueous (Bq L^{-1}) and solid phases (Bq kg^{-1}), resulting in the same final units as above, i.e., L kg^{-1}. A large K_d value is indicative of a greater tendency for a solute to partition to the solid phase; thus, a greater degree of infiltration and leaching will be required for the contaminant to move through the soil and vadose system into the underlying groundwater, and then migrate down gradient.

Several relevant databases have been developed for using the linear partitioning function to describe radionuclide transport [64, 81, 88–92]. The US Environmental Protection Agency (EPA) in collaboration with DOE issued a series of comprehensive reports discussing the underlying geochemical processes controlling contaminant sorption, including several key radionuclides (i.e., Am, Cs, ^{3}H, I, Np, Pu, Ra, Rn, Sr, Tc, Th, U), and the often-cited limitations associated with using the linear partitioning coefficient (i.e., K_d) to describe contaminant migration in the subsurface environment. A major recommendation noted throughout the reports is the necessity to derive site-specific K_d values to better understand a given contaminant transport scenario. The use of literature-derived K_d values should be restricted to preliminary screening calculations of contaminant transport when additional data is not readily available. EPA did, however, produce a concise table listing important aqueous and solid-phase parameters that impacted the sorption behavior of each inorganic contaminant discussed in the reviews (see Table 2) [81, 89, 90]. It is interesting to note that field-derived K_d values are often much higher than lab estimates because extended exposure tends to reduce the apparent contaminant solubility, often described as aging.

In 2010, the International Atomic Energy Agency (IAEA) released an extensive compendium of estimated partition coefficients that could be used in evaluating the impact of both routine discharges and accidental releases. The IAEA report covers a more extensive list of radionuclides than the previous EPA reports, and further distinguishes transfer parameters based on soil properties, such as texture and organic matter content (Table 3). The classification of soil texture is limited to three main groups (sand, loam and clay), compared to 12 textural classes in the US Department of Agriculture system. A fourth class defined as organic soils is reserved for soils with >20% organic matter content. The IAEA report also distinguishes K_d values based on additional key soil properties, termed cofactors, that control radionuclide partitioning, in many respects consistent to the parameters identified by the EPA in Table 2.

Radionuclide Fate and Transport in Terrestrial Environments. Table 2 Important aqueous- and solid-phase parameters that influence radionuclide partitioning in the soil environment [89, 90]

Element	Important aqueous- and solid-phase parameters influencing contaminant sorption
Am	Clay content, Fe/Al oxide content, pH
Cs	Clay content, Fe/Al oxide content, (NH_4^+), CEC, mica/illite clay content, pH, (K^+)
I	Dissolved halides, OM content, pH, redox, volatilization
Pu	Clay content, Fe/Al oxide content, (CO_3^{2-}, F^-, SO_4^{2-}, PO_4^{3-}), OM content, pH, redox
Ra	Clay content, Fe/Al oxide content, $BaSO_4$ coprecipitation, (dissolved alkaline earth elements), CEC, ionic strength, OM content, pH
Rn	None
Sr	CEC, (Ca), (CO_3^{2-}), pH, (Sr)
Tc	OM content, redox
Th	Fe/Al oxide content, (CO_3^{2-}), OM content, pH
^{3}H	None
U	Clay content, Fe/Al oxide content, (CO_3^{2-}, F^-, SO_4^{2-}, PO_4^{3-}), OM content, pH, redox, (U)

CEC cation exchange capacity, *OM* organic matter
Parentheses represent aqueous concentrations

In addition to the Linear partitioning coefficient (K_d), the Freundlich and Langmuir isotherms are widely used in describing solute partitioning in the soils. For comparison, an example data set with the three common isotherms is provided in Fig. 3. The Freundlich isotherm is

$$S = K_F C^n \tag{20}$$

where S and C are the same as defined above, K_F is the Freundlich partition coefficient (L kg^{-1}), and n is the dimensionless heterogeneity parameter that describes the nonlinearity of the sorption process (n is typically ≤ 1). Note that when $n = 1$ the Freundlich equation is identical to the Linear isotherm. A plot of log S versus log C should yield a straight line with a slope of n and an intercept of log K_F. Although it describes the reduction in contaminant sorption often observed with increasing concentration, the Freundlich equation fails to account for the sorption maximum that is often observed.

Originally used to describe the adsorption from gases to solids, the Langmuir isotherm is

$$S = \frac{S_m K_L C}{1 + K_L C} \tag{21}$$

where S_m is the maximum adsorption per unit mass of sorbent, K_L is the Langmuir affinity parameter, and, as above, C is the solute concentration in equilibrium (Mol L^{-1}, mg L^{-1}, etc.) with the sorbent phase. Values for K_L and S_m are typically derived from batch sorption data using various linear transformations of the Langmuir equation or nonlinear least squares error fitting routines [58].

Radionuclide Fate and Transport in Terrestrial Environments. Table 3 Partition coefficients (K_d) for radionuclides as a function of soil texture [64, 107]

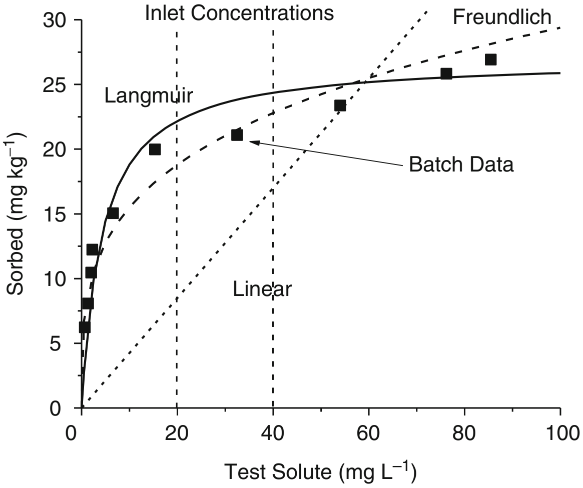
Element	Soil group	N^a	Mean^b	GSD^c	Minimum	Maximum
Sr	All soils	255	5.2×10^{1}	5.9	4.0×10^{-1}	6.5×10^{3}
	Sand	65	2.2×10^{1}	6.4	4.0×10^{-1}	2.4×10^{3}
	Loam + clay + organic	176	6.9×10^{1}	5.4	2.0	6.5×10^{3}
Cs	All soils	469	1.2×10^{3}	7.0	4.3	3.8×10^{5}
	Sand	114	5.3×10^{2}	5.8	9.6	3.5×10^{4}
	Loam + clay	227	3.7×10^{2}	3.6	3.9×10^{-1}	3.8×10^{5}
	Organic	108	2.7×10^{2}	6.8	4.3	9.5×10^{4}
U	All soils	178	2.0×10^{2}	12	7.0×10^{-1}	6.7×10^{4}
	Mineral	146	1.8×10^{2}	13	7.0×10^{-1}	6.7×10^{4}
	Organic	9	1.2×10^{3}	6.1	3.3×10^{2}	7.6×10^{3}
Th	All soils	46	1.9×10^{3}	10	1.8×10^{1}	2.5×10^{5}
	Mineral	25	2.6×10^{3}	10	3.5×10^{1}	2.5×10^{5}
	Organic	5	7.3×10^{2}	44	1.8×10^{1}	8.0×10^{4}
I	All soils	250	6.9	5.4	1.0×10^{-2}	5.8×10^{2}
	Mineral	196	7.0	5.2	1.0×10^{-2}	5.4×10^{2}
	Organic	11	3.2×10^{1}	3.3	8.5	5.8×10^{2}

[a] Number of data sets
[b] Geometric mean
[c] Geometric standard deviation

As apparent in Fig. 3, both the Freundlich and Langmuir isotherms provide a good description of the entire data set, accounting for an initial high level of sorption that decreases with increasing concentration. In this instance, the slope of the Linear partitioning coefficient was chosen to intersect with the two other isotherms. The Linear isotherm could have more accurately described initial solute partitioning at relatively low surface coverages. However, the choice of an arbitrary solution concentration (or surface coverage) for evaluating K_d is a drawback to the Linear partitioning approach. Other inherent limitations of the empirical sorption isotherms will become more apparent when discussing solute transport.

Mechanistic Adsorption Models: Surface Complexation Modeling

Surface complexation models (SCM) address some of the inherent uncertainties in the empirical partitioning (K_d) approach by describing sorption in terms of surface functional groups, with formation constants, reaction stoichiometries, and in some cases sorption site densities determined by fitting batch sorption data, or exhaustive physicochemical characterization of the sorptive material. Several approaches to SCM have been developed (e.g., Constant Capacitance Model, Double Layer Model, Triple Layer Model, etc.), generally differing in the manner in which they describe the mineral/water interface, and account for columbic interactions that impact the mass action equations. SCM models generally conform to the four following principles: (1) reactive surfaces are composed of surface functional groups that react with dissolved solutes to form surface species in a manner similar to complexation reactions that occur in solution; (2) surface acidity reactions that generate charge can be described by SCM, and the resulting surface charge can be invoked based on electrical double-layer theory; (3) the apparent mass action binding constants describing sorption



Radionuclide Fate and Transport in Terrestrial Environments. Figure 3
Example batch sorption data described using the three most common empirical sorption isotherms, Linear, Freundlich, and Langmuir. The dashed lines at 20 and 40 mg L^{-1} reflect the two inlet column concentrations used in generating the breakthrough curves presented in Fig. 4

reactions are empirical parameters that can be related to thermodynamic constants through the use of activity coefficients for the surface species; and (4) the electrical charge at the mineral surface is determined by the reactions of surface function groups [58, 86, 93–95].

As discussed previously for amphoteric surface functionalities, surface charge and electrical potential in SCM are derived from the consequences of surface complexation for protons, hydroxyls, and other sorbates, with specific surface reactive sites defined as *MOH* (Eq. 7), where M represents the metal ion of the oxide mineral (i.e., Fe, Al, etc.) that is bound to the reactive surface hydroxyl. For phyllosilicate clays, the edges form the sites for surface complexation reactions through aluminol and silanol groups [57]. Protonation and deprotonation reactions are described as follows:

$$MOH + H^+ \leftrightarrow MOH_2^+ \tag{22}$$

$$MOH \leftrightarrow MO^- + H^+ \tag{23}$$

Reaction 22 and 23 are somewhat simplistic in that potentiometric titrations of oxide minerals often fail to display clear inflection points because surface acidity is a function of surface protonation and site heterogeneity even in fairly ideal monomineralic systems [86]. One approach for addressing such site heterogeneity is the use of strong and weak coordination sites, as in the Diffuse Double Layer Model (DDLM) of Dzombak and Morel (1990, ref. [96]). Even so, the thermodynamic equilibrium for Eq. 23 can be written as:

$$K = \frac{[MO^-](H^+)\gamma_{\mathrm{MO}}}{[MOH]\gamma_{\mathrm{MOH}}} \tag{24}$$

where the brackets refer to concentration, and parenthesis refer to activity, and γ is the activity coefficient for the respective species. Instead of defining more sites, the consequences of surface site heterogeneity discussed above are often addressed within the activity corrections in order to limit the number of site types that are modeled. Equilibrium constants derived for surface complexation reactions, such as the one described in Eq. 24, differ from aqueous reactions in that the surface activity terms represent corrections for the electrostatic energy at the mineral solution interface, i.e., the electrical double layer (*edl*):

$$K = K_C \times edl \tag{25}$$

where K_C is a conditional equilibrium constant written in terms of the activities for aqueous species, concentrations for surface sites, and the *edl* term is a model-dependent means of accounting for the surface potential or potential within the adsorptive plane [97]. One limitation in applying SCM to natural systems is the difficulty of characterizing the electrical double-layer properties of heterogeneous soils and aquifer sediments where sorbed ions, organic matter, and oxide coatings alter mineral surface charge properties, i.e., $\mathrm{pH}_{\mathrm{pzc}}$, in a manner that is difficult to predict and evaluate experimentally [77, 86, 87, 93, 98]. Goldberg et al. (2007) provide an excellent review of the experimental techniques used in validating a mechanistic understanding of contaminant sorption processes [58].

Two distinct approaches to the application of SCM to describe contaminant sorption to soil and aquifer materials have developed: the Component Additivity (CA) approach and the Generalized Composite (GC) approach. In the CA approach, one assumes the sorption behavior of heterogeneous natural materials can be described using a relative combination of the individual reactive components, with specific mass law equations accounting for each phase of the mixture based on sorption experiments conducted on individual synthetic or ideal mineral analogs. For highly idealized systems, the CA approach can be quite effective in describing sorption behavior and even multicomponent solute transport [99]. For example, Table 4 provides the surface species and stability constants for UO_2^{2+}sorption to the poorly crystalline Fe oxyhydroxide, i.e., ferrihydrite, utilizing weak and strong binding sites [95, 100]. Although defining "strong" and "weak" sites enhances the ability to account for surface heterogeneity, it also doubles the number of adjustable site parameters, i.e., protonation/deprotonation constants, contaminant complexation constants, etc. However, the application of the CA approach to describing sorption behavior for natural materials has achieved only limited success [58, 101].

In contrast, the GC approach assumes that the sorption behavior of natural materials is far too complex to be described using the contributions of each sorptive phase. Furthermore, GC is often invoked in a non-electrostatic manner that considers adsorption purely in terms of surface equilibria without correcting for electrostatics, the influence of which is included to some degree in the resulting binding constants. As such, the resulting model may favor surface reactions that lack stoichiometric or spectroscopic validity, while providing an accurate description of the macroscopic sorption behavior. When the semiempirical GC approach is invoked, the specific protonation of the reactive surfaces does not have to be explicitly included, as it is grossly included within the contaminant specific sorption reactions [58, 93, 98, 101].

Vegetation Transfer Factors

Similar to the soil partition coefficient (K_d), the soil-to-plant transfer factor (TF), sometimes referred to as the Concentration Ratio (CR), is often used to quantify the bioavailability or plant uptake of a given soil contaminant. More mechanistic approaches to predicting plant accumulation of soil radionuclides based on nutrient uptake models have been developed. However, the inherent complexity of the rhizosphere and difficulty in parameterizing such models have limited their application [102].

The TF is the concentration of the radionuclide in dried plant tissue (Bq kg^{-1}) divided by the concentration of the radionuclide in the dried soil (Bq kg^{-1}) supporting the plant of interest, where the units cancel out yielding a dimensionless transfer term. Since mobility is a key factor controlling plant uptake of radionuclides, the TF is inversely correlated with K_d as expected [103]; however, the degree of correlation is often insufficient for predictive purposes. The TF is largely assumed to be associated with root uptake, as resuspension and foliar deposition are generally assumed to be minor [64, 104]. However, contaminated

Radionuclide Fate and Transport in Terrestrial Environments. Table 4 Stability constants for U(VI) sorption to Fe oxides (ferrihydrite) in component additivity (CA) modeling approach [95, 100]

Surface Species	Log K
$TOH + H^+ \leftrightarrow TOH_2^+$	6.51
$WOH + H^+ \leftrightarrow WOH_2^+$	6.51
$TOH \leftrightarrow TO^- + H^+$	−9.13
$WOH \leftrightarrow WO^- + H^+$	−9.13
$T(OH)_2 + UO_2^{2+} \leftrightarrow TO_2UO_2^O + 2H^+$	−2.57
$W(OH)_2 + UO_2^{2+} \leftrightarrow WO_2UO_2^O + 2H^+$	−6.28
$T(OH)_2 + UO_2^{2+} + H_2CO_3 \leftrightarrow TO_2UO_2CO_3^{2-} + 4H^+$	−13.031
$W(OH)_2 + UO_2^{2+} + H_2CO_3 \leftrightarrow WO_2UO_2CO_3^{2-} + 4H^+$	−17.103

T and *W* refer to strong and weak binding sites, respectively.

particulates are often a major part of the initial source term and risk associated with past nuclear reactor accidents [62, 105].

Although similar to the K_d, the TF depends on numerous factors, including the initial source of contamination (i.e., fallout, liquid waste, etc.), the duration since introduction, specific characteristics of the radionuclide, soil properties, soil management practices, nutrient status, and plant species, to name a few [104, 106, 107]. Microbial processes independent of plants also influence speciation, mobility, and plant availability of soil-borne radionuclides, particularly through the control of soil redox status. As observed for essential trace nutrients, mycorrhizal fungi can also facilitate root uptake of radionuclides, such as ^{137}Cs and ^{90}Sr [108, 109]. In addition, the calculated TF will depend to a large degree on how the rooting zone is defined for the purposes of soil sampling and analysis. The International Union of Radioecologists (IUR) has established a standard for estimating TF, with the upper 10 cm chosen as the rooting depth for grass species, and the upper 20 cm for all other crops, including fruit trees [106, 110]. The 2010 IAEA report also provides uptake transfer factors for 14 plant groups reflecting a range of crop species. The reader is directed to IAEA (2009) for a more comprehensive discussion of the complete data set and methods used in deriving the IAEA (2010) transfer parameters [64, 107].

Transfer factors tend to decrease with time after source introduction as the bioavailability of the radionuclide decreases. In recognition that soil properties, organic matter turnover, and nutrient cycling differ in response to the predominant environmental conditions, the IAEA [64] further divided the TF data into tropical, subtropical, and temperate systems based on site climate. As an example, an abbreviated list of the TF values for ^{90}Sr^{2+} associated with cereal grain production under various climatic conditions is presented in Table 5. As evident in Table 5, far less data are available for the subtropical and tropical systems in general. Furthermore, a specific set of TF values have been derived for rice production because of its importance as major food source in the humid tropics, and the unique flooded conditions under which it is grown. The fluctuating redox conditions associated with rice production can impact the speciation of redox-sensitive radionuclides, such as U, Pu, I, Tc, as well as the stability and sorption capacity of Fe

Radionuclide Fate and Transport in Terrestrial Environments. Table 5 Transfer factor (TF) values for ^{90}Sr^{2+} in cereal crops under various climatic conditions [64, 107]

Temperate								
Element	Plant group	Plant compartment	Soil group	N	Mean/value	GSD/SD	Minimum	Maximum
Sr	Cereals	Grain	All	282	1.1×10^{-1}	2.7	3.6×10^{-3}	1.0
			Sand	123	1.4×10^{-1}	3.0	3.6×10^{-3}	1.0
			Loam	71	1.1×10^{-1}	2.4	1.6×10^{-2}	7.2×10^{-1}
			Clay	72	7.8×10^{-2}	2.4	5.3×10^{-3}	7.1×10^{-1}
			Organic	10	9.7×10^{-2}	4.1	1.2×10^{-2}	3.6×10^{-1}
Subtropical								
Sr	Cereals	Grain	Loam	8	5.1×10^{-2}	1.3	3.6×10^{-2}	6.5×10^{-2}
		Stems, shoots	Loam	7	1.5×10^{-1}	2.5	4.2×10^{-2}	4.2×10^{-1}
Tropical								
Sr	Cereals	Grain	All	2	6.0×10^{-1}		4.4×10^{-1}	7.6×10^{-1}
			Sand	1	4.4×10^{-1}			
			Loam	1	7.6×10^{-1}			

N number of entries, *GSD/SD* geometric standard deviation/standard deviation oxyhydroxides like goethite (α-FeOOH) and ferrihydrite ($Fe(OH)_3$), important sorbents in soil systems.

Radionuclide Migration

To date, contaminant transport research and modeling efforts have generally focused on improving the understanding of water movement while much less attention has been placed on the interrelated biological and geochemical reactions that impact solute partitioning, treating such complex processes in a rather simplistic manner that can be easily implemented within a multidimensional flow model [59, 111, 112]. However, numerous studies have demonstrated the inherent limitations in applying simplistic empirical partitioning expressions to describing contaminant migration at the field scale. Although coupling various reactive partitioning mechanisms to a transport model can be numerically challenging, a major limitation to the implementation of more mechanistic solute partitioning in transport models is the defensible parameterization of the model [59, 113].

For simplicity, radionuclide partitioning and transport in soil will be discussed in terms of the one-dimensional, steady-state convection dispersion equation (CDE):

$$\begin{array}{l} \dfrac{\partial C}{\partial t} = D\dfrac{\partial^2 C}{\partial z} - v\dfrac{\partial C}{\partial z} - \dfrac{\rho}{\theta}\dfrac{\partial S}{\partial t} \\ \text{Terms} \qquad 1 \qquad\quad 2 \qquad\quad 3 \end{array} \tag{26}$$

where C is the concentration of the solute in the aqueous phase (M L^{-3}), S is the sorbed concentration per unit mass of solid phase (M M^{-1}), ρ is the porous media bulk density (M L^{-3}), θ is the volumetric water content (L^3 L^{-3}), v is the flow velocity (L T^{-1}), t or T is time (sec., min., days, etc.), z is distance (L), and D is the dispersion coefficient (L^2 T^{-1}). Starting from the left, the CDE indicates that changes in the concentration of a solute with time are a reflection of the hydrodynamic dispersion (term 1), linear convective flow velocity (term 2), and sorption (term 3). In some form, the CDE has been widely applied to the description of radionuclide migration in soils, e.g., [64, 114–116]. Additional terms can be added to the equation to account for multiple flow domains, biodegradation, first-order decay, and other related processes. Also, preferential flow, physical and chemical soil heterogeneity, fluctuating θ, and/or facilitated transport may account for the rapid migration that is often observed for a small fraction of radionuclides following initial release to the soil environment [61, 115, 117].

The dispersion coefficient (L^2 T^{-1}) can be further defined as:

$$D = \lambda v + D_e \tag{27}$$

where λ is the dispersivity (L), and D_e is the effective diffusion coefficient (L^2 T^{-1}). A non-sorbing solute, also referred to as a conservative tracer, is transported convectively in the flow stream and spreads by dispersion, a combination of molecular diffusion and mechanical dispersion resulting from differences in pore-water velocity and flow path. Thus, dispersion serves to dilute the solute plume (i.e., radionuclide) over the course of migration.

Understanding the physics of water movement is essential to correctly describing the migration behavior of a reactive solute, and interpreting the underlying mechanisms [118–120]. Dispersivity, generally considered a scale-dependent function of the physical heterogeneity encountered during solute migration, is a key factor in describing the physical aspects of contaminant migration in the subsurface environment. In addition to transport scale, different λ values have been reported as a function of varying degrees of saturation, θ. The transmissive path for convective transport is altered by changes in the degree of saturation that may control reaction kinetics and the type of surfaces available for sorption (i.e., phyllosilicate clays, Fe oxides, etc.) [120, 121]. Extensive breakthrough tailing observed under any flow regime can result from both physical and/or chemical disequilibria that may be conveniently, although incorrectly, described by the dispersion coefficient or even multiple flow domains, especially when applying a transport model that lacks the ability to explicitly account for nonlinear sorption behavior [122–124].

Assuming one understands the flow regime, linear sorption (i.e., n = 1) can be described as $S = K_d C$ (Eq. 19), and the CDE equation (Eq. 26) can be further simplified:

$$R\frac{\partial C}{\partial t} = D\frac{\partial^2 C}{\partial z} - v\frac{\partial C}{\partial z} \tag{28}$$

where R represents the dimensionless retardation factor as it relates to K_d, ρ_{BD} is the matrix bulk density (M L^{-3}), and the volumetric water content, θ:

$$R = 1 + \frac{\rho_{BD} K_d}{\theta} \tag{29}$$

Retardation of a given solute can be more generally defined as

$$R = \frac{v_w}{v_s} \tag{30}$$

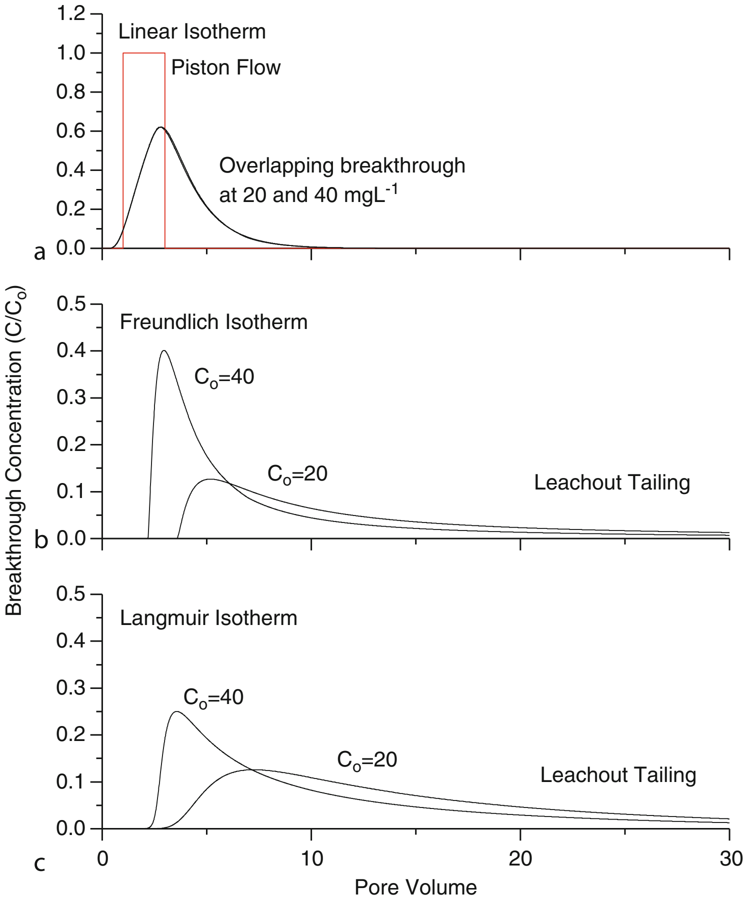
where v_w is the travel velocity for water, and v_s is the travel velocity of the solute of interest. A value other than 1 indicates that the solutes interact with the soil in some way, with $R > 1$ indicative of sorption and $R < 1$ indicating some form of pore-water exclusion that results in the solute traveling faster than water.

Differences in K_d values observed for batch equilibration and flow-through systems have been attributed to: (1) differences in soil:solution ratio inherent to each system; (2) failure to remove or account for competitive reaction products in batch studies; (3) limitations in the availability of sorption sites in batch versus column experiments; (4) colloid-facilitated transport; (5) inherent limitations in the linear distribution coefficient to account for temporal and spatial variations in groundwater chemistry; (6) partitioning reactions controlled by solubility and/or redox conditions; and (7) rate-limited geochemical reactions and physical nonequilibrium present under flowing versus static conditions [62, 120, 125–130]. Under complex field conditions, several of these limitations may be operative at the same time. However, flow-through techniques can provide a better representation of field conditions than batch sorption experiments by accounting for temporal and spatial nonequilibrium.

When radionuclide sorption can be described using the K_d, the CDE equation (Eq. 28) can be solved analytically given appropriate boundary conditions, such as pulse-like or step input histories. More complex reaction schemes and variable boundary conditions can be invoked within the transport equation through numerical solutions. Figure 4 illustrates the breakthrough behavior observed for the three different sorption isotherms presented in Fig. 3, i.e., Linear, Freundlich, and Langmuir. In discussing the modeled breakthrough results, the term “rad” will be used to identify the hypothetical radionuclide solute. In each case, the rad is introduced at one of two different inlet concentrations (20 and 40 mg L^{-1}) to illustrate the impact of solute concentration on breakthrough for the three different sorption isotherms. For both cases the inlet rad concentration (C_o) remains constant for two pore volumes (PV), followed by several PV of rad-free water. In Fig. 4a, piston flow behavior for a nonreactive solute is included for comparison to illustrate the influence of both sorption and dispersion on contaminant transport. For piston flow, the solute displays 100% breakthrough (i.e., $C/C_o = 1$) after exactly 1 PV of leaching, and the leach-out breakthrough occurs exactly 1 PV after the inlet source is removed, i.e., 3 PV in the present example. As a result of sorption and hydrodynamic dispersion, the reactive solute breakthrough is both delayed and diluted with respect to the piston flow example, with maximum breakthrough of about 60% ($C/C_o = 0.6$) occurring at slightly less than 3 PV. However, the breakthrough behavior is exactly the same for the two different inlet concentrations, and the total rad mass applied to the column is recovered in the effluent.

Similar breakthrough is observed for both the Freundlich (Fig. 4b) and Langmuir (Fig. 4c) isotherms at each of the two different inlet concentrations, with both earlier and higher relative breakthroughs observed with increasing inlet concentration. Differences between the two nonlinear isotherms would be more evident at higher inlet concentrations where the Langmuir isotherm has leveled off at the sorption maximum, while the Freundlich isotherm continues to predict increasing sorption. Also, extensive breakthrough tailing observed for both nonlinear isotherms is a consequence of the higher relative degree of sorption observed at lower concentrations when compared to the Linear isotherm. Despite such different breakthrough histories, the total rad mass applied is recovered in the column effluent for all three isotherms, assuming sufficient leaching.

As illustrated by the previous example, the interpretation of contaminant migration behavior suffers from the same limitations inherent in the interpretation of sorption behavior in laboratory batch studies. Even in relatively simple one-dimensional column studies, similar contaminant migration patterns can be derived from different combinations of chemical



Radionuclide Fate and Transport in Terrestrial Environments. Figure 4
Relative outlet concentration (i.e., C/Co) as a function of time (pore volume) for a reactive solute (2 PV pulse) based on sorption isotherm: (**a**) Linear, (**b**) Freundlich, and (**c**) Langmuir partitioning (based on isotherms in Fig. 3)

retardation processes (e.g., reaction kinetics, nonlinear sorption behavior, sorption competition, etc.) and physical transport phenomena, such as mass transfer kinetics [118, 120, 126, 131, 132]. At a fundamental level, the mathematical equations used to describe physical and chemical nonequilibrium are identical, and thus distinguishing such processes based on the transport behavior of a reactive solute requires the presence of a "conservative" tracer for comparison. Reactive solutes often display complex transport behavior that can be misinterpreted in terms of multiple chemical and physical processes in the absence of an a priori understanding of the solute sorption behavior [122, 124, 133–136].

A transient flow model that accounts for changes in soil θ, usually through the use of Richards' equation to describe water flow, can improve transport predictions for radionuclides displaying low to moderate partitioning to the immobile soil matrix [137, 138]. For radionuclides displaying very high K_d values, movement in the soil environment may be insensitive to transient fluctuations in θ, and thus migration can be described by the steady-state CDE based on the average rate of water movement down the profile. As before, the dispersion term accounts for a combination of mechanisms that tend to spread the solute plume in a manner analogous to diffusion. Furthermore, various forms of the CDE equation can be used to describe radionuclide distribution in the soil profile without considering contaminant partitioning based on "apparent" convection and dispersion coefficients that describe the center of mass and distribution, respectively, without accounting for actual water flux or θ. This approach yields an apparent site-specific migration rate (i.e., cm y^{-1}) and an apparent dispersion coefficient that can be readily used for comparing the behavior of various radionuclides [115, 116, 139–141]. The application of the steady-state CDE model to unsaturated soil systems often under predicts an initial period of rapid migration when applied to the transport of radionuclides with high K_d values, generally assumed to reflect physical or chemical nonequilibrium.

Compartmental models, also called "residence time" models, offer a more "black box" alternative to the CDE for describing radionuclide migration in the soil and vadose zone. The soil profile is split into "compartments" that are conceptually consistent with natural soil horizonization, providing a means for accounting for vertical soil heterogeneity. However, model compartment dimensions are often established based on an arbitrary sampling scheme rather than soil properties.

Radionuclide transport in a compartment model is controlled by a system of linear first-order differential equations representing each layer [139, 142]. The change in concentration (i.e., activity) of a radionuclide, C, in a given layer, i, can be described as:

$$\frac{\Delta C_i}{\Delta t} = K_{i-1}C_{i-1} - K_iC_i \tag{31}$$

where K_i is the transfer rate from one layer to another. The resulting migration velocity for the compartmental model reflects a combination of both convective and diffusive transport processes, yielding a migration rate that differs from the effective convection velocity in the CDE. In the compartmental model, the dispersive effect is a function of layer thickness in that the accuracy of radionuclide distribution is limited to the thickness of layers homogenized for soil analysis [115, 139, 143, 144]. As with the CDE, both analytical and numerical methods are available depending on the complexity of both the compartmental layers and the underlying boundary conditions.

Although the SCM approach has been used extensively to describe batch sorption data [96], extension of the approach to the description of reactive solute migration has been limited. To date, the SCM approach has been reasonably successful in describing U, zinc (Zn), phosphate ($PO_4{}^{3-}$), and molybdate (MoO_4^{2-}) migration in saturated materials, and boron (B) in an irrigated agricultural field [93, 98, 101, 145–148]. As discussed above for even relatively simple solute transport models, the development of mechanistic transport models must be constrained in a manner to minimize the number of model parameters that can lose their mechanistic significance when applied to the complexities of the field.

Characteristics of Specific Radionuclides

In discussing the properties of radionuclides that impact their behavior in soils, it is important to note that the physical and chemical processes that control solubility, sorption, and migration have no impact on the isotopes decay mechanism or radioactivity. Information about the total amount of a given radionuclide in a soil system provides little information concerning its form, i.e., chemical speciation, or bioavailability [16]. Furthermore, radionuclides do represent a unique hazard when compared to other environmental contaminants (i.e., heavy metals, organic solvents, etc.) in that intimate contact, inhalation, and ingestion are not necessarily prerequisites for exposure. Formed by neutron capture reactions, actinides (e.g., Pu, Am, and Np) exhibit complex speciation chemistry with multiple stable oxidation states, each of which can form complexes and solid phases with unique stability characteristics [13, 82]. Because of such complexity, a thorough discussion of actinide speciation chemistry is beyond the scope of the current chapter.

Americium

Americium, an actinide with no stable isotopes, has mainly been introduced in the environment through atmospheric weapons testing in the 1950s and 1960s. Nineteen isotopes of Am have been identified with atomic masses ranging from 231 to 249, most of which have very short half-lives. Americium-241 ($t_{1/2}$ = 432.2 y) and ^{243}Am ($t_{1/2}$ = 7,370 y) are the only isotopes with half-lives long enough to persist as radioactive waste in SNF. Plutonium-241 ($t_{1/2}$ = 14.4 y) results from multiple neutron adsorption events for ^{238}U, which then decays via β emission to form ^{241}Am. Americium-241 primarily decays by α emission to ^{237}Np plus a strong γ emission that is useful as a source for γ radiography. Americium is also used as a source of ionization in smoke detectors [90, 149].

Similar to several other actinides (Pa, U, Np, Pu), Am can occur in several oxidation states (III–VI), although the +3 form dominates with a speciation chemistry similar to trivalent lanthanides. All other oxidation states are strong oxidizing agents and are therefore unstable in the presence of other oxidizable compounds. As with other actinides, Am forms strong complexes with oxygen-donating ligands (OH^-, PO_4^{3-}, and CO_3^{2-}) and relatively weak complexes with monovalent anions, Cl^-. The Am^{3+} cation dominates under acidic conditions, with various hydrolysis products forming with increasing pH. It is usually considered to be fairly immobile, displaying relatively high K_d values; however, this can make facilitated transport more likely. For instance, Artinger et al. [150] found that a small fraction of Am was preferentially transported in association with DOC, while the largest fraction remained sorbed to immobile Fe oxides. Similar to U, the Am-carbonate complexes are generally more important with increasing pH, forming anionic complexes that increase solubility under alkaline conditions. Americium also forms strong complexes with humic substances [12, 50, 82, 90, 149].

In a relevant example of Am partitioning in soils, leaking storage drums at the Rocky Flats DOE site released ^{241}Am and ^{239,240}Pu. Both Am and Pu were found to be largely associated with Fe oxides and to lesser degree with SOM. Column and batch experiments indicated that reducing conditions in the presence of strong complexing agents were effective at recovering a large fraction of the total α activity from the Rocky Flats soil [151]. Another study at Rocky Flats measured ^{239,240}Pu and ^{241}Am in storm-water runoff and pond-water discharge in different size fractions [50]. The actinides were found to be primarily associated with the particulate phase (>0.5 μm), but a significant colloidal fraction was able to pass through the filter.

Cesium

Cesium, with ^{133}Cs as the only stable isotope, is ubiquitous in the soil environment and exists in the +1 oxidation state, forming very few stable aqueous complexes. The most important radioactive Cs^+ isotopes include: ^{134}Cs ($t_{1/2} = 2.05$ y), ^{135}Cs ($t_{1/2} = 3 \times 10^6$ y), and ^{137}Cs ($t_{1/2} = 30.2$ y). Cesium-137, a nuclear fission product, decays by β^- emission with ≈ 85% forming the metastable ^{137m}Ba which subsequently decays within minutes by γ emission (0.66 MeV) [10, 12].

As a chemical surrogate for K^+, Cs^+ behavior is governed by similar metabolic and geochemical processes in the environment. Like K^+, Cs^+ tends to serve as a biochemical electrolyte rather than a structural component in plant or animal tissues. However, Cs^+ concentrations and Cs^+/K^+ ratios have been observed to increase by a factor of 2–3 for each increase in the trophic level. Despite similar uptake rates, the apparent bioaccumulation of Cs^+ has been attributed to the longer biological retention times when compared to K^+, with a biological half-life in humans that is two to five times that of K^+ [152].

Despite its high solubility, Cs^+ can be selectively adsorbed and strongly bound to zeolites and certain 2:1 phyllosilicate clays, often resulting in contamination being restricted in close proximity to the point of deposition [61, 153–157]. As discussed earlier, polyvalent ions tend to sorb to a greater degree than monovalent ions on soil clays; however, there is a distinct exception when discussing monovalent ions with low hydration energy, such as Cs^+, Rb^+, and K^+ (also $NH_4{}^+$). Micas (illites) and vermiculites, 2:1 phyllosilicate clays possessing relatively high layer charge, have the ability to irreversibly fix or "sorb" certain elements and molecules within the clay interlayer, significantly reducing their mobility and bioavailability in the soil environment.

Interlayer sorption followed by cation dehydration can induce interlayer collapse known as fixation, which is to some degree irreversible [153, 158–160]. Such high-affinity interlayer sorption sites are called frayed edge sites (FES selectivity $Cs^+ > Rb^+ > NH_4{}^+ > K^+$) [161, 162]. The specific Cs^+ sorption capacity for soil and clay materials has been operationally defined by use of the silver thiourea (Ag-TU) method. The Ag-TU complex displays a high affinity for cation exchange sites on the outer planar surfaces of soil clays, blocking their ability to retain common soil cations; however, the Ag-TU complex is unable to access "frayed edge sites" because of its size [163]. Other factors such as the soil moisture content, the presence of organic acids and cations of similar ionic radius (K^+ and $NH_4{}^+$) can affect Cs^+ sorption and fixation by soil clays [158, 164–166].

This suggests that the relative effectiveness interlayer fixation of Cs^+ as a stabilization method depends on factors such as background solution composition (i.e. predominant cation), soil pH, clay mineralogy, and SOM content [166].

Recognizing the widespread distribution of 2:1 clays and their ability to strongly sorb ^{137}Cs, the National Council on Radiological Protection and Measurement (NCRP) recommends a default value for the TF of 0.1 for risk calculations. However, TF values reported for the DOE Savannah River Site (SRS) facility, located near Aiken, SC, are much higher, with mean values ranging from $\approx$ 3 to 15 [167–169]. The poor retention in SRS soils has been attributed to their coarse texture and a clay fraction dominated by Fe oxides (i.e., goethite) and kaolinite, a 1:1 layer phyllosilicate with no capacity for interlayer fixation [73]. The poor retention of Cs may be exacerbated in riparian and wetland systems where anoxic conditions can result in the buildup of NH_4^+ that competes with Cs^+ for FES sorption sites [167, 170, 171].

In systems that display a limited intrinsic ability to fix $^{137}Cs^+$, the addition of illitic clays can effectively reduce the mobility and bioavailability of $^{137}Cs^+$ [73]. Using a limno-corral system that isolates the water column above the contaminated pond sediments, Hinton et al. [170, 172] further demonstrated that the addition of illites can reduce $^{137}Cs^+$ availability in aquatic systems. However, the widespread use of such a remediation technique is limited by the need to incorporate the sorptive materials within the contaminated soils, a practice that would drastically disturb a functioning, albeit contaminated, ecosystem.

Greenhouse and field studies have demonstrated the ability of zeolites to reduce the uptake of $^{137}Cs^+$ in plants [41, 173]. Certain zeolites, e.g., clinoptilolite, display high ion exchange selectivities for Cs^+ and Sr^{2+} [41], even in the presence of Ca^{2+} and Mg^{2+}, and are naturally abundant enough for use as low-cost sorbents. Selectivity for Cs^+ and Sr^{2+} is generally highest under slightly acidic conditions (pH 5.0–6.0), but drops off dramatically at lower pH values due to competition with hydrogen for exchange sites. The presence of strong ligands, such as the synthetic chelate EDTA or citrate and tartate, can also reduce Cs^+ and Sr^{2+} sorption [174]. One advantage of zeolites is that they possess high sorption capacity even in larger particle-size fractions due to the internal reactive voids. This makes the material ideal for systems where hydraulic conductivity must be maintained, such as reactive flow-through barriers. Zeolites can also be incorporated into bentonite clay liners in minor amounts to improve the sorptive capacity of the liner without overly enhancing hydraulic conductivity. However, the larger particle size can have a negative impact on the kinetics of sorption [174]. In addition, zeolites are not effective sorbents for transuranic species, such as uranyl (UO_2^{2+}), that are commonly found at sites with elevated Cs^+ and Sr^{2+} levels [39].

Iodine

Iodine (I) has many short-lived isotopes (including ^{131}I and ^{125}I), one long-lived isotope, ^{129}I ($t_{1/2}$ = 1.6 million y), and one stable isotope, ^{127}I. Due to its long half-life, tendency to accumulate in the thyroid, and perceived high mobility in the environment, ^{129}I is often a key radionuclide of environmental concern [175]. There are many sources of ^{129}I to the environment. Naturally, ^{129}I can be produced from cosmic-ray spallation of Xe in the atmosphere and spontaneous fission of ^{238}U in the earth's crust. Anthropogenic inputs of ^{129}I to the environment include atmospheric weapons testing, the Chernobyl reactor accident, and SNF reprocessing, with the major contributors being Sellafield (formerly Windscale) and LaHague [176].

Iodine occurs in several oxidation states with −1, +5, and molecular I_2 being the most environmentally relevant [81]. Iodide (I^-), iodate (IO_3^-) and organic iodine are the main species of iodine in the environment. Although I^- and IO_3^- dominate in environmental waters, organic iodine is common in soils and sediments [177]. Iodine forms negatively charged species and thus, like ^{99}Tc, is assumed to be highly mobile in neutral to alkaline environments. This behavior is of particular concern at contaminated sites such as the DOE-SRS. The F-Area Basins on the SRS were used for the disposal of large volumes of low-level acidic waste. After basin closure, remediation efforts first involved pump-and-treat to control the spread of the mobile groundwater contaminants, i.e., ^3H_2O, ^{129}I, etc. More recently, base injection to neutralize the low pH (~3.5) groundwater has been successfully applied to sequester contaminant metals; however, such treatment can mobilize ^{129}I associated with variable-charge minerals as a result of the increase in pH. Current studies are underway to better understand I speciation and to help develop successful remediation strategies [178–180].

Historically, total ^{129}I has been measured by accelerator mass spectrometry and neutron activation. More recently, analytical methods have involved the use of high performance liquid chromatography and inductively coupled plasma mass spectrometry coupled with ion chromatography and gas chromatograph mass spectrometry to better evaluate the dynamics of I speciation [177, 180–182].

Plutonium, Pu

The current environmental inventory of Pu is largely derived from nuclear weapons testing and accidental releases, such as the reactor fire at Windscale in 1957 and the Chernobyl explosion in 1986 [10, 28]. However, minor amounts of Pu were generated from neutron capture by ^{238}U within the U ore of the Oklo natural reactor in Gabon. There are 15 known Pu isotopes, the three most common of which are ^{239}Pu (94%; $t_{1/2}$ = 24,000 y), ^{240}Pu (6%; $t_{1/2}$ = 6,560 y), and ^{241}Pu (0.4%; $t_{1/2}$ = 14.4 y). Plutonium-239 and ^{240}Pu decay by α emission, forming ^{235}U and ^{236}U, respectively. The generation of fissile ^{239}Pu from the more plentiful ^{238}U associated with nuclear fuel contributes about one-third of the energy produced in the reactor. Plutonium is a major proliferation hazard because a limited mass of ^{239}Pu ($\approx$ 7–10 kg) is required to create a nuclear bomb. When ingested, Pu tends to concentrate in the bones and liver [10, 12].

Oxidation-reduction reactions are critical in controlling Pu behavior in the environment, with four common redox states (i.e., III, IV, V, and VI) possible under typical pH and redox conditions. In fact, multiple Pu redox states can coexist in solution. Iron oxides such as goethite (α-FeOOH), hematite (α-Fe_2O_3), and magnetite (Fe_3O_4) have been shown to facilitate the reduction from Pu(V) to Pu(IV) across a broad pH range, from 3 to 8, with sorption to the oxide required in many instances for reduction to occur [183, 184]. Plutonium (IV) is less soluble and sorbs readily. Plutonium forms very strong carbonate and hydroxyl complexes that enhance solubility at pH > 7 [21, 82, 89, 185, 186].

Radium-226 and Radon-222

There are two products within the U-Th decay series that warrant special attention, Radium-226 (^{226}Ra, $t_{1/2}$ = 1,600 y) and its daughter Radon-222 (^{222}Rn, $t_{1/2}$ = 3.8 d), because they represent a significant fraction of the background internal radiation dose to humans. Radium (Ra^{2+}), an alkali-earth metal, behaves similarly to Ca^{2+} and Ba^{2+}, and is often found associated with U-bearing minerals. Radium may coprecipitate with various sulfate and carbonate minerals, and readily sorbs to phyllosilicate clays and Fe and Al oxides in a fairly reversible manner. Radium is considered to be more mobile in the soil environment than U and more subject to plant uptake [12, 59].

Radon (Rn) is a colorless and odorless inert gas that is water soluble, insensitive to aqueous speciation, and displays negligible interaction (i.e., $K_d \approx 0$) with soil components. However, the generation of ^{222}Rn depends on the presence of the parent elements (i.e., U, Th, Ra) that are subject to various soil processes, i.e., sorption, precipitation, etc. Radon-222 ($t_{1/2}$ = 3.8 d) is derived from ^{226}Ra via decay in the ^{238}U decay series. Twenty isotopes of Rn have been identified, all of them radioactive, many with very short half-lives [89]. The inert ^{222}Rn gas displays a negligible ability to accumulate in the body; however, the short half-life results in the significant potential for α decay to occur while exposed to the sensitive tissues within the lungs, accounting for approximately two-thirds of the natural radiation dose within the USA. Furthermore, Rn exposure has been estimated to account for 6–12% of the cases of lung cancer observed in the USA [10].

Strontium-90

Strontium (Sr), a chemical analog for Ca^{2+}, is generally present in the soil environment in the +2 oxidation state. Nuclear fuel processing and weapons testing are the primary sources for the radioactive fission products, $^{89}Sr^{2+}$ ($t_{1/2}$ = 52 d) and $^{90}Sr^{2+}$ ($t_{1/2}$ = 29 y). Because of the short half-life of $^{89}Sr^{2+}$, $^{90}Sr^{2+}$ represents the more persistent environmental problem. As with other alkaline earth cations, Sr^{2+} partitioning is primarily controlled by the CEC of the soil, and competition with other common exchangeable cations (i.e., Ca^{2+}, Mg^{2+}, Na^+, K^+, etc.), as well as stable Sr^{2+}. Solution pH and ionic strength are also important factors controlling Sr^{2+} partitioning [187–189].

On an equivalence basis, cation exchange experiments have shown that stable Sr^{2+} will dominate cation exchange site in competition with alkaline and alkaline earth cations, i.e., $Sr^{2+} > Ca^{2+} > Mg^{2+} > K^{+} > NH_4^{+} > Na^{+}$; however, the concentration of $^{90}Sr^{2+}$ in soil systems is typically several orders of magnitude lower than Ca^{2+}. Consistent with cation exchange as a sorption mechanism, $^{90}Sr^{2+}$ does not become more recalcitrant with extended equilibration time, and is therefore subject to desorption when in the presence of competing cations [190]. However, other studies have reported that some fraction of $^{90}Sr^{2+}$ may become less exchangeable with extended soil equilibration, presumably due to substitution within calcium-bearing minerals. Under high pH conditions, $^{90}Sr^{2+}$ may coprecipitate with calcite ($CaCO_3$) or anhydrite ($CaSO_4$), or precipitate as strontianite, $SrCO_3$ [89, 191]. Strontium does not form strong complexes with either organic or inorganic ligands, with the possible exception of $Sr\text{-}CO_{3(aq)}$ under high pH conditions [192]. As an analog of Ca^{2+}, Sr^{2+} is readily taken up by plants and tends to accumulate in the bone and bone marrow of animals [12].

Technetium-99

Technetium-99 (^{99}Tc), with a half-life of 210,000 years, is a neutron fission product of ^{235}U and ^{239}Pu [12]. Under oxidizing conditions, Tc persists in the environment as the pertechnetate anion (TcO_4^{-}), with Tc in the +7 oxidation state. As an anion, pertechnetate is generally assumed to sorb poorly to many soil clays that display significant negative surface charge (i.e., cation exchange capacity) resulting from isomorphic substitution [12, 193]. Technetium-99 also appears to be readily translocated in plants from the roots to leaves.

Pertechnetate (TcO_4^{-}) sorption is generally associated with variable-charge soil (i.e., amphoteric) minerals, like Fe and Al oxides/oxyhydroxides. However, TcO_4^{-} is subject to competition for sorption sites from other common soil anions, such as Cl^{-}, NO_3^{-}, PO_4^{3-}, and SO_4^{2-}. Sheppard and Thibault (1990) reported Tc (VI) K_d values in the range from 0.1 to 1.0 depending on texture and SOM content [92]. In contrast, Kaplan et al. (2008) reported higher K_d values for materials rich in Fe oxides [194]. Under anoxic conditions, Tc (VII) can be reduced to the sparingly soluble Tc(IV) oxidation state, which in the absence of strong complexing agents is subject to additional sorption to soil minerals and organic matter and may form the sparingly soluble $TcO_2{\cdot}nH_2O$ [195–197]. Heterogeneous reduction of Tc(VI) by sorbed Fe(II) is faster than homogeneous reduction by soluble Fe(II). Such reactions are somewhat reversible as recent studies have shown that Fe(III) can serve as an oxidant for Tc(IV) under acidic condition [196].

Thorium

Thorium (Th) is ≈ three times more abundant in the earth than U, and exists in the +4 oxidation state, forming hydroxide and carbonate species in the soil environment. Thorium has several isotopes with atomic masses ranging from 212 to 236, all of which are unstable [10, 12, 89]. In addition to its greater abundance than U, interest in the use of Th for nuclear power generation has increased in recent years because it represents less of a nuclear proliferation threat, and produces fewer long-lived actinides (i.e., Np, Am, and Cs). For use in a nuclear reactor, ^{232}Th requires a neutron source to produce the fissile ^{233}U; however, ^{232}Th can replace natural uranium (i.e., mixture of 0.7% ^{235}U and 99.3% ^{238}U) in the reactor fuel with less enriched U required as an initial neutron source. Once the process has started, Th-based fuel has the potential to maintain (i.e., breed) the production of ^{233}U without requiring any additional enriched U. Further, the use of a fuel mixture of Th and natural U reduces the production of ^{239}Pu and complicates the potential extraction of fissile ^{233}U, further limiting nuclear proliferation when compared to current fuel practices [17].

Tritium, ^{3}H

Tritium (^{3}H) is one of the three isotopes of hydrogen, which include the stable isotopes hydrogen (^{1}H; 99.984% abundance) and deuterium (^{2}H; 0.016% abundance). As an isotope of hydrogen, ^{3}H can exist in ionic, gaseous, solid, and liquid forms. Small quantities of ^{3}H are produced naturally in the stratosphere by cosmic-ray interaction with ^{14}N to form ^{12}C in addition to ^{3}H. Nuclear fallout and discharges from nuclear power facilities reflect a greater source of ^{3}H

than cosmic rays. Tritium is produced by fission and thermal neutron reactions with ^{6}Li in nuclear reactors. Tritium has a half-life of 12.3 y, and readily combines with oxygen to form tritiated water (^3H_2O), with a $K_d \approx 0$. Tritium decays to ^{3}He by means of low-energy β^- emission (18.6 keV) that is commonly determined by liquid scintillation analysis. Atmospheric nuclear weapons testing has released 2.4 × 10^{20} Bq of ^{3}H into the environment [10].

Tritium has several industrial uses including luminous paints, non-powered light sources, and as a radioactive tracer in chemical, geological, and biological experiments [12]. Because ^3H_2O water travels at the same rate as groundwater (i.e., $K_d \approx 0$ and R = 1), it is often used as a conservative tracer in natural recharge and solute transport experiments to differentiate between multiple complex physical and chemical processes that impact the migration of a "reactive" solute [88, 188]. Although subject to possible substitution for stable H on oxides and clays, such reactions are deemed to be minor because of the relative low concentration of ^{3}H compared to the other isotopes of hydrogen, ^{1}H and ^{2}H, indicating that ^{3}H would need to be highly preferred to see significant retention, as indicated by the lack of an effective ion exchange method for separating $^3H^+$.

Tritiated water can be incorporated into the body by several pathways, including inhalation of vapor, absorption through the skin, and ingestion, by far the dominant uptake pathway [198, 199]. Once inside the body, ^3H_2O is generally distributed in the bodily fluids, with a limited fraction incorporated into various body compartments, and predominant excretion through urine and feces. A small fraction of ingested ^{3}H is tightly bound to organic structures, i.e., organically bound tritium (OBT), which can be released by decomposition or combustion, complicating liquid scintillation analysis [200]. With continuous exposure, ^3H_2O reaches a steady-state concentration in body fluids, and removal of the source allows for depuration [198, 201, 202]. The human biological half-life for ingested ^3H_2O within body fluids ranges from about 9.5 to 12 days compared to ≈ 2.26 days for rodents [12, 203, 204].

Uranium

For a comprehensive discussion of U geochemistry the readers are directed to Burns and Kinch [205]. Uranium is the most common radionuclide soil contaminant throughout the DOE weapons complex [23]. Furthermore, thermodynamically unstable $UO_{2(s)}$ represents greater than 90% of SNF, which upon oxidation forms $UO_{3(s)}$ corrosion products at relevant environmental temperatures [206]. In nature, U exists primarily as three isotopes, ^{238}U (99.3% abundance; $t_{1/2} = 4.5 \times 10^9$ y), ^{235}U (0.71%; $t_{1/2} = 7.13 \times 10^8$ y), and ^{234}U (0.0057%; $t_{1/2} = 2.47 \times 10^5$ y). However, ^{235}U is of primary interest for nuclear weapons and power generation because of its fissile capacity [10, 12]. Uranium exists in multiple oxidation states (from +4 to +6), with U(IV) and U(VI) dominating under reducing and oxidizing conditions, respectively. Under oxidizing conditions, U(VI) hydrolyzes instantly to form the uranyl cation (UO_2^{2+}). In the absence of carbonates, UO_2^{2+} forms various hydrolysis species in response to pH, UO_2OH^+, $UO_2(OH)_2^o$, $UO_2(OH)_2^-$, etc. [95]. Soil parameters such as OM content, Fe oxide content, and pH play a strong role in controlling U partitioning [207].

The U-carbonate species tend to dominate at neutral and higher pH conditions. Uranium also forms strong complexes with fluoride (F^-), phosphate (PO_4^{3-}), sulfate (SO_4^{2-}), and DOC [205, 208–211]. Cation exchange studies suggest that UO_2^{2+} displays similar affinity to the fully hydrated Ca^{2+} for interlayer cation exchange sites; however, surface complexation reactions tend to dominate with increasing U concentrations and pH > 4 [212]. In a biological setting, U tends to accumulate in the bones and bone marrow of animals, with a biological half-life of ≈ 300 d. Long-lived radionuclides, like natural and depleted U (i.e., mostly ^{238}U), may display greater chemical rather than radiological toxicity [12, 213].

Most remediation efforts focus on altering U speciation to favor the formation of insoluble/immobile U phases. Amonette et al. [214] provide an in-depth review of various methods for assessing the environmental availability of U as a means of evaluating the relative effectiveness of such efforts. Uranium reduction in situ has been widely evaluated as a potential remediation strategy [215–219]. In recognition of the fact that many actinides and other heavy metals form relatively insoluble phosphates under a range of pH conditions, the addition of apatite ($Ca_5(PO_4)_3OH$) and other PO_4 sources has been evaluated as a means of remediating U and other metal-contaminated sites [73, 74, 84, 120, 211, 220–224]. For example, Arey et al. (1998) demonstrated in batch studies that minor additions (<5% by wt.) of hydroxyapatite to a highly weathered contaminated soil from the DOE-SRS significantly reduced water-soluble U, as well as a number of other environmentally important metals, such as nickel (Ni) and lead (Pb). Although directly correlated to DOC, U solubility in apatite-amended soil was found to be less than that expected for the common U phosphate mineral, autunite, with TEM analysis revealing U to be largely associated with secondary Al-phosphate precipitates rather than residual apatite [74, 84].

Future Directions

In concluding the discussion of processes controlling the fate and transport of radionuclides in terrestrial systems, it is important to note critical areas warranting greater research attention. As mentioned previously, a good deal of site-specific data is required for the development of both empirical and mechanistic models (both conceptual and numerical) for describing radionuclide partitioning in the presence and absence of corrective measures, and their impact on receptor exposure. The development of efficient geophysical methods for evaluating physical and chemical heterogeneity that impact radionuclide partitioning at the field scale is critical in site characterization and identifying an appropriate remediation strategy, including MNA and institutional control. Such characterization efforts should be a component of a systematic, iterative process designed to minimize uncertainty associated with performance assessment and waste disposition.

Additional research is needed concerning the aqueous and solid-phase speciation of transuranics in complex environmental samples, including aqueous complex formation and their incorporation at trace levels within mineral phases. Assessing the impact of facilitated transport (i.e., mobile colloids and DOC) for strongly sorbing/insoluble radionuclides requires additional consideration. Further development and refinement of a consistent thermodynamic database for radionuclides that can be used within mechanistic speciation, transport, and exposure models will aid in reducing uncertainties associated with environmental impact assessments. Even when the use of mechanistic partitioning schemes within contaminant transport models is impractical, a better understanding of the partitioning mechanisms can certainly help to constrain the use of appropriate empirical parameters for a given transport scenario. Continued focus on reaction and partitioning kinetics is warranted as many environmental systems are not at chemical equilibrium. An expanded focus on the ecological impact of radionuclides in combination with other anthropogenic contaminants (i.e., metals, organic solvents, etc.) is warranted, as contaminated sites usually contain a complex mixture of various hazardous agents as part of a multi-stressor approach.

Acknowledgments

Preparation of this manuscript was partially funded by Cooperative Agreement DE-FC09-07SR22506 between the University of Georgia and the Department of Energy. The authors would like to acknowledge the assistance of T. Murphy and L. Lee with revising the manuscript and reproducing the figures.

Abbreviations

AO	Ammonium-oxalate
AEC	Anion exchange capacity
amu	Atomic mass units
CEC	Cation exchange capacity
CDB	Citrate-dithionite-bicarbonate
CA	Component additivity
CDE	Convection dispersion equation
DOE	Department of Energy
DDLM	Diffuse Double Layer Model
EPA	Environmental Protection Agency
FES	Frayed edge sites
GC	Generalized composite
HLNWR	High-level nuclear waste repository
HIS	Hydroxyl-interlayered smectite
HIV	Hydroxyl-interlayered vermiculite
IAEA	International Atomic Energy Agency
IUR	International Union of Radioecologists
IAP	Ion activity product
MCL	Maximum contaminant limit
MNA	Monitored natural attenuation
NCRP	National Council on Radiological Protection and Measurement

NORM	Naturally occurring radioactive materials
NRC	Nuclear Regulatory Commission
OBT	Organically bound tritium
OM	Organic matter
PV	Pore volume
SI	Saturation index
SRS	Savannah River Site
SEM	Scanning electron microscope
Ag-TU	Silver thiourea
SOM	Soil organic matter
TF	Soil-to-plant transfer factor
SNF	Spent nuclear fuel
SCM	Surface complexation modeling
TCLP	Toxicity Characteristic Leaching Procedure
UMTRA	Uranium Mill Tailings Remedial Action
XRD	X-ray diffraction

Bibliography

1. Strand P, Brown J (2009) The international conference on radioecology and environmental radioactivity. J Environ Rad 100:999–1001
2. Fuma S, Ishii N, Takeda H, Miyamoto K, Yanagisawa K, Doi K, Kawaguchi I, Tanaka N, Inamori Y, Polikarpov GG (2009) Effects of acute γ-irradiation on the aquatic microbial microcosm in comparison with chemicals. J Environ Rad 100:1027–1033
3. Pennington CW (2007) Exposing America: Comparative assessments of ionizing radiation doses to U.S. populations from nuclear and non-nuclear industries. Prog Nucl Energy 49:473–85
4. Abdelouas A (2006) Uranium mill tailings: Geochemistry, mineralogy and environmental impact. Elements 2(6):335–341
5. APS (2007) Consolidated interim storage of commercial spent nuclear fuel- A technical and programmatic assessment, Am Phys Soc. 19
6. NRC (2008) Review of DOE's nuclear energy research and development program, Committee on review of DOE's nuclear energy research and development program, National Research Council. p 102
7. Hu QH, Weng JQ, Wang JS (2010) Sources of anthropogenic radionuclides in the environment: A review. J Environ Rad 101:426–437
8. Peters MT, Ewing RC, Steefel CI (2008) GNEP waste form campaign science & technology and modeling & simulation program: Roadmap with rationale & recommendations, Global Nuclear Energy Partnership. p 51
9. Lieser KH (2001) Nuclear and radiochemistry: Fundamentals and applications, 2nd edn. Wiley-VCH, Weinheim, Federal Republic of Germany, 462
10. Loveland W, Morrissey DJ, Seaborg GT (2006) Modern nuclear chemistry. Wiley-Interscience, Inc, Hoboken, NJ
11. Rydberg J, Liljenzin JO, Choppin G (2001) Radiochemistry and nuclear chemistry. Elsevier Science, Woburn, MA, p 720
12. Zhang PC, Krumhansl JL, Brady PV (2002) Introduction to properties, sources and characteristics of soil radionuclides. In: Zhang PC, Brady PV (eds) Geochemistry of soil radionuclides. Soil Science Society of America, Madison, WI, pp 1–20
13. Langmuir D (1997) Aqueous environmental geochemistry. Prentice Hall, Upper Saddle River, NJ, p 600
14. Ivanovich M (1982) The phenomenon of radioactivity. In: Ivanovich M, Harmon RS (eds) Uranium series disequilibrium: Applications to environmental problems. Clarendon Press, Oxford, pp 1–32
15. Friedlander G, Kennedy JW, Macias ES, Miller JM (1981) Nuclear and radiochemistry, 3rd edn. John Wiley & Sons, New York, p 684
16. Elless MP, Lee SY (2002) Radionuclide-contaminated soils: A mineralogical perspective for their remediation. In: Dixon JD, Schulze DG (eds) Soil mineralogy with environmental applications. Soil Science Society of America, Madison, WI, pp 737–763
17. Macfarlane AM, Miller M (2007) Nuclear energy and uranium resources. Elements 3:185–192
18. Linsley G (2001) International advice and experience relevant to chronic radiation exposure situations in the environment. In: Brechignac F, Howard BJ (eds) Radioactive pollutants: Impact on the environment. EDP Sciences, France, pp 105–129
19. IAEA (2003) Extent of environmental contamination by naturally occurring radioactive materials (NORM) and technological options for remediation. International Atomic Energy Agency, Vienna
20. Aarkrog A (2001) Thoughts on radioecology by the millennium shift. In: Brechignac F, Howard BJ (eds) Radioactive pollutants: Impact on the environment. EDP Sciences, France, pp IX–XXI
21. Choppin GR (2003) Actinide speciation in the environment. Radiochim Acta 91:645–649
22. Ewing RC (2006) The nuclear fuel cycle: A role for mineralogy and geochemistry. Elements 2(6):331–34
23. Riley RG, Zachara JM, Wobber FJ (1992) Chemical contaminants on DOE lands and selection of contaminant mixtures for subsurface science research. DOE, Washington, DC, p 77
24. Glagolenko YV, Drozhko YG, Rovny SI (2009) Experience in rehabilitating contaminated land and bodies of water around the Mayak production association. In: Schweitzer GE, Parker FL, Robbins K (eds) Cleaning up sites contaminated with radioactive materials. NRC, Washington, DC, pp 81–91
25. Garwin RL (2001) Can the world do without nuclear power? Can the world live with nuclear power? Interdiscip Sci Rev 26(4):265–271
26. McBride JP, Moore RE, Whitherspoon JP, Blanco RE (1978) Radiological impact of airborne effluents of coal and nuclear plants. Science 8:1045–1050
27. Kharecha PA, Kutscher CF, Hansen JE, Mazria E (2010) Options for near-term phaseout of CO_2 emissions from coal in the United States. Environ Sci Technol 44(11):4050–4062
28. Jones S (2007) Windscale and Kyshtym: A double anniversary. J Environ Radioact 99:1–6

29. IAEA (2005) Chernobyl's legacy: Health, environmental and socio-economic impacts and recommendations to the governments of Belarus, the Russian federation and Ukraine. International Atomic Energy Agency, Vienna
30. Klein C, Hurlbut CS (1993) Manual of mineralogy, 21st edn. John Wiley & Sons, New York, NY
31. Dixon JB, Schulze DG (eds) (2002) Soil mineralogy with environmental applications. Soil Science Society of America, Madison, WI, p 866
32. Harris W, White GN (2008) X-ray diffraction techniques for soil mineral identification. In: Methods of soil analysis Part 5–Mineralogical methods. Soil Science Society of America, Madison, WI, pp 81–115
33. Fourdrin C, Allard T, Monnet I, Menguy N, Benedetti M, Calas G (2010) Effect of radiation-induced amorphization on smectite dissolution. Environ Sci Technol 44:2509–2514
34. Barnhisel RI, Bertsch PM (1989) Chlorites and hydroxy-interlayered vermiculite and smectite. In: Minerals of the soil environment. Soil Science Society of America, Madison, WI, pp 729–788
35. Breck DW (1974) Zeolite molecular sieves. Krieger Publishing Company, Malabar, FL
36. Sparks DL, Huang PM (1985) Physical chemistry of soil potassium. In: Munson RD (ed) Potassium in agriculture. American Society of Agronomy, Madison, WI, pp 201–276
37. Ming DW, Mumpton FA (1989) Zeolites in soils Ch. 18. In: Dixon JB, Weed SB (eds) Minerals in the soil environment. Soil Science Society of America, Madison, WI, pp 973–997
38. Dyer A (2000) Application of natural zeolites in the treatment of nuclear wastes and fall-out. In: Cotter-Howells JD, Campbell LS, Valsami-Jones E, Batchelder M (eds) Environmental mineralogy: Microbial interactions, anthropogenic influences, contaminated land, and waste management. The Mineralogical Society of Great Britain and Ireland Oxford, UK
39. Vaniman DT, Bish DL (1995) The importance of zeolite in the potential high-level radioactive waste repository at Yucca Mountain, nevada. In: Ming DW, Mumptom FA (eds) Natural zeolites, '93 occurrence, properties, use. International committee on natural zeolites, Brockport, NY, pp 533–546
40. Boettinger JL, Ming DW (2002) Zeolites. In: Dixon JB, Schulze DG (eds) Soil mineralogy with environmental applications. Soil Science Society of America, Madison, WI, pp 585–610
41. Chelishchev NF (1995) Use of natural zeolite at Chernobyl. In: Ming DW, Mumptom FA (eds) Natural zeolites, '93 occurrence, properties, use. International committee on natural zeolites Brockport, NY, pp 525–532
42. Schulze DG (2002) An Introduction to soil mineralogy: Chapter 1. In: Dixon JB, Schulze DG (eds) Soil mineralogy with environmental applications. Soil Science Society of America, Madison, WI, pp 1–35
43. Schwertmann U, Taylor RM (1989) Iron oxides. In: Dixon JB, Weeds SB (eds) Minerals in the soil environment. Soil Science Society of America, Madison, WI, pp 379–438
44. Jackson ML, Lin CH, Zelazny LW (1986) Oxides, hydroxides, and aluminosilicates. In: Klute A (ed) Methods of soil analysis, vol 1, Physical and mineralogical methods. American Society of Agronomy Inc, Madison, WI, pp 101–150
45. Gerth J (1990) Unit-cell dimensions of pure and trace metal-associated goethites. Geochim Cosmochim Acta 54: 363–371
46. Eary LE, Rai D (1987) Kinetics of chromium (III) oxidation to chromium (VI) by reaction with manganese dioxide. Environ Sci Technol 21:1187–1193
47. Manceau A, Charlet L, Boisset MC, Didier B, Spadini L (1992) Sorption and speciation of heavy metals on hydrous Fe and Mn oxides. From microscopic to macroscopic. Appl Clay Sci 7:201–223
48. Kinniburgh DG, vanRiemsdijk W, Koopal L, Borkovec M, Benedetti M, Avena M (1999) Ion binding to natural organic matter: Competition, heterogeneity, stoichiometry and thermodynamic consistency. Colloids Surf, A 151:147–66
49. Deng Y, Dixon JB (2002) Soil organic matter and organic-mineral interactions. In: Dixon JB, Schulze DG (eds) Soil mineralogy with environmental applications. Soil Science Society of America, Madison, WI, pp 69–107
50. Santschi PH, Roberts KA, Guo LD (2002) Organic nature of colloidal actinides transported in surface water environments. Environ Sci Technol 36(17):3711–3719
51. McCarthy JF, Zachara JM (1989) Subsurface transport of contaminants. Environ Sci Technol 23:496–502
52. Kersting AB, Efurd DW, Finnegan DL, Rokop DJ, Smith DK, Thompson JL (1999) Migration of plutonium in groundwater at the Nevada test site. Nature 397:56–59
53. Zeh P, Czerwinski KR, Kim JI (1997) Speciation of uranium in Gorleben groundwaters. Radiochim Acta 76:37–44
54. Dearlove JPL, Longworth G, Ivanovich M, Kim JI, Delakowitz B, Zeh P (1991) A study of ground-water colloids and their geochemical interactions with natural radionuclides in gorleben aquifer systems. Radiochim Acta 52/53:83–89
55. Xu C, Santschi PH, Zhong JY, Hatcher PG, Francis AJ, Dodge CJ, Roberts KA, Hung C-C, Honeyman BD (2008) Colloidal cutin-like substance cross-linked to siderophore decomposition products mobilizing plutonium from contaminated soils. Environ Sci Technol 42(22):8211–8217
56. Sposito G (1984) The surface chemistry of soils. Oxford University Press, New York, 234
57. Goldberg S (2005) Chapter 10: Equations and models describing adsorption processes in soils. In: Tabatabai MA, Sparks DL (eds) Chemical processes in soils. Soil Science Society of America, Madison, WI, pp 489–517
58. Goldberg S, Criscenti LJ, Turner DR, Davis JA, Cantrell KJ (2007) Adsorption-desorption processes in subsurface reactive transport modeling. Vadose Zone J 6:407–435
59. IAEA (2006) Applicability of monitored natural attenuation at radioactively contaminated sites. International Atomic Energy Agency, Vienna, p 105
60. Brady PV, Jove-Colon CF, Carr G, Huang F (2002) Soil radionuclide plumes. In: Zhang PC, Brady PV (eds) Geochemistry of soil radionuclides. Soil Science Society of America, Madison, WI, pp 165–190

61. Alexakhin RM, Krouglov SV (2001) Soil as the main compartment for radioactive substances in terrestrial ecosystems. In: Brechignac F, Howard BJ (eds) Radioactive pollutants: Impact on the environment. EDP Sciences, France, pp 149–174
62. Salbu B (2009) Challenges in radioecology. J Environ Radioact 100:1086–1091
63. Varga B, Leclerc E, Zagyvai P (2009) The role of analogues in radioecology. J Environ Rad 100:802–805
64. IAEA (2010) Handbook of parameter values for the prediction of radionuclide transfer in terrestrial and freshwater environments. International Atomic Energy Agency, Vienna, p 194
65. Runde W, Neu MP, Conradson SD, Li D, Lin M, Smith DM, Van-Pelt CE, Xu Y (2002) Geochemical speciation of actinides in soil and solution. In: Zhang PC, Brady PV (eds) Geochemistry of soil radionuclides. Soil Science Society of America, Madison, WI, pp 45–59
66. Seaman JC, Guerin M, Jackson BP, Bertsch PM, Ranville JF (2003) Analytical techniques for characterizing complex mineral assemblages: Mobile soil and groundwater colloids. In: Selim HM, Kingery WL (eds) Geochemical and hydrological reactivity of heavy metals in soils. Lewis Publishers, New York, pp 271–309
67. McCarthy JF, Degueldre C (1993) Sampling and characterization of colloids and particles in groundwater for studying their role in contaminant transport: Chapter 6. In: Buffle J, Leeuwen HPv (eds) Environmental particles. Lewis Publishers, Ann Arbor, MI
68. Tessier A, Campbell PGC, Bisson M (1979) Sequential extraction procedure for the speciation of particulate trace metals. Anal Chem 51:844–850
69. Miller WP, Martens DC, Zelazny LW (1986) Effect of sequence in extraction of trace metals from soils. Soil Sci Soc Am J 50:598–601
70. Clark SB, Johnson WH, Malek MA, Serkiz SM, Hinton TG (1996) A comparison of sequential extraction techniques to estimate geochemical controls on the mobility of fission product, actinide, and heavy metal contaminants in soils. Radiochim Acta 74:173–179
71. Rigol A, Roig M, Vidal M, Rauret G (1999) Sequential extractions for the study of radiocesium and radiostrontium dynamics in mineral and organic soils from western europe and Chernobyl areas. Environ Sci Technol 33(6):887–895
72. Voegelin A, Barmettler K, Kretzschmar R (2003) Heavy metal release from contaminated soils: Comparison of column leaching and batch extraction results. J Environ Qual 32:865–875
73. Seaman JC, Meehan T, Bertsch PM (2001) Immobilization of ^{137}Cs and U in contaminated sediments using soil amendments. J Environ Qual 30(4):1206–1213
74. Arey JS, Seaman JC, Bertsch PM (1999) Immobilization of uranium in contaminated sediments by hydroxyapatite addition. Environ Sci Technol 33:337–342
75. Evangelou VP, Phillips RE (2005) Cation exchange in soils. In: Chemical processes in soils. Soil Science Society of America, Madison, WI, pp 343–410
76. Stumm W, Morgan JJ (1995) Aquatic chemistry: An introduction emphasizing chemical equilibria in natural waters, 3rd edn. Wiley-Interscience, New York
77. Bertsch PM, Seaman JC (1999) Characterization of complex mineral assemblages: Implications for contaminant transport and environmental remediation. P Natl Acad Sci 96:3350–3357
78. Grenthe I, Fuger J, Konings RJM, Lemire RJ, Muller AB, Nguyen-Trung C, Wanner H (1992) Chemical thermodynamics of uranium. North Holland, Amsterdam
79. Lovley DR, Phillips EJP, Gorby YA, Landa ER (1991) Microbial reduction of uranium. Nature 350(April 4):413–416
80. Gorby YA, Lovley D (1992) Enzymatic uranium precipitation. Environ Sci Technol 26:205–207
81. USEPA (1999) Understanding variation in partition coefficient, K_d, values- Volume I: The K_d model, measurement, and application of chemical reaction codes. US Environmental Protection Agency, Washington, DC
82. Runde W (2002) Geochemical interactions of actinides in the environment. In: Zhang PC, Brady PV (eds) Geochemistry of soil radionuclides. Soil Science Society of America, Madison, WI, pp 21–44
83. Karathanasis AD (2002) Mineral equilibria in environmental soil systems. In: Soil mineralogy with environmental applications. Soil Science Society of America, Madison, WI, pp 109–151
84. Seaman JC, Arey JS, Bertsch PM (2001) Immobilization of Ni and other metals in contaminated sediments using soil amendments. J Environ Qual 30(2):460–469
85. Zachara JM, Smith SC, Resch CT, Cowan CE (1992) Cadmium sorption to soil separates containing layer silicates and iron and aluminum oxides. Soil Sci Soc Am J 56:1074–1084
86. Davis JA, Kent DB (1990) Surface complexation modeling in aqueous geochemistry. Rev Mineralog 23:177–260
87. Coston JA, Fuller CC, Davis JA (1995) Pb^{2+} and Zn^{2+} adsorption by a natural aluminum- and iron-bearing surface coating on an aquifer sand. Geochim Cosmochim Acta 59(17):3535–3547
88. Thibault DH, Sheppard MI, Smith PA (1990) A critical compilation and review of default soil solid/liquid partition coefficients, Kd, for use in environmental assessments. Atom Energy Canada Ltd AECL10125:1–112
89. USEPA (1999) Understanding variation in partition coefficient, K_d, values- Volume II: Review of geochemistry and available Kd values for cadmium, cesium, chromium, lead, plutonium, radon, strontium, thorium, tritium, and uranium. US Environmental Protection Agency, Washington, DC
90. USEPA (2004) Understanding variation in Partition coefficient, K_d, values Volume III: Review of geochemistry and available Kd values for americium, arsenic, curium, iodine, neptunium, radium, and technetium, Office of Air and Radiation, Washington, DC
91. Looney BB, Grant MW, King CM (1987) Estimation of geochemical parameters for assessing subsurface transport at the Savannah River Site. E.I. du Pont de Nemours & Co, Aiken, SC
92. Sheppard MI, Thibault DH (1990) Default soil solid/liquid partition coefficients, K_dS, for four major soil types: A compendium. Health Phys 59:471–482

93. Davis JA, Coston JA, Kent DB, Fuller CC (1998) Application of the surface complexation concept to complex mineral assemblages. Environ Sci Technol 32(19):2820–2828
94. Sposito G (1983) On the surface complexation model of the oxide-aqueous solution interface. J Colloid Interface Sci 91:329–340
95. Davis JA, Payne TE, Waite TD (2002) Simulating the pH and pCO_2 dependence of uranium(VI) adsorption by a weathered schist with surface complexation models. In: Zhang PC, Brady PV (eds) Geochemistry of soil radionuclides. Soil Science Society of America, Madison, WI, pp 61–86
96. Dzombak DA, Morel FMM (1990) Surface complexation modeling: Hydrous ferric oxide. John Wiley & Sons, Inc, New York, p 393
97. Zachara JM, Westall JC (1998) Chemical modeling of ion adsorption in soils. In: Sparks DL (ed) Soil physical chemistry. CRC Press, Boca Raton, FL, pp 47–95
98. Davis JA, Randall JD (2003) Application of surface complexation modeling to describe uranium (VI) adsorption and retardation at the uranium mill tailings site at Naturita, Colorado. U. S. Nuclear Regulatory Commission Office of Nuclear Regulatory Research, Washington, DC, pp 1–223
99. Meeussen JCL, Scheidegger A, Hiemstra T, Riemsdijk WHV, Borkovec M (1996) Predicting multicomponent adsorption and transport of fluoride at variable pH in a goethite-silica sand system. Environ Sci Technol 30:481–488
100. Waite TD, Davis JA, Payne TE, Waychunas GA, Xu N (1994) Uranium (VI) adsorption to ferrihydrite: Application of a surface complexation model. Geochim Cosmochim Acta 58(24):5465–5478
101. Davis JA, Meece DE, Kohler M, Curtis GP (2004) Approaches to surface complexation modeling of uranium(VI) adsorption on aquifer sediments. Geochim Cosmochim Acta 68(18): 3621–3641
102. Casadesus J, Sauras T, Gonze MA, Vallejo R, Brechignac F (2001) A nutrient-based mechanistic model for predicting the root uptake of radionuclides. In: Brechignac F, Howard BJ (ed) Radioactive pollutants: Impact on the environment. EDP Sciences, France, pp 209–239
103. Sheppard SC, Sheppard MI (1989) Impact of correlations on stochastic estimates of soil contamination and plant uptake. Health Phys 57:653–657
104. Vandenhove H, Olyslaegers G, Sanzharova N, Shubina O, Reed E, Shang Z, Velasco H (2009) Proposal for new best estimates of the soil-to-plant transfer factor of U, Th, Ra, Pb and Po. J Environ Radioact 100:721–732
105. Salbu B, Lind OC, Borretzen P, Oughton DH (2001) Advanced speciation techniques for radionuclides associated with colloids and particles. In: Brechignac F, Howard BJ (ed) Radioactive pollutants: Impact on the environment. EDP Sciences, France, pp 243–260
106. IAEA (2000) Radiation legacy of the 20th century: Environmental restoration, in RADLEG 2000. International Atomic Energy Agency, Moscow, Russian Federation, 64 p
107. IAEA (2009) Quantification of radionuclide transfers in terrestrial and freshwater environments for radiological assessments. International Atomic Energy Agency, Vienna p 616
108. Tamponnet C, Plassard C, Parekh N, Sanchez A (2001) Impact of micro-organisms on the fate of radionuclides in rhizospheric soils. In: Brechignac F, Howard BJ (eds) Radioactive pollutants: Impact on the environment. EDP Sciences, France, pp 175–185
109. Strandberg M, Johansson M (1998) ^{134}Cs in heather seed plants grown with and without mycorrhiza. J Environ Rad 40(2):175–184
110. IUR (1992) Protocol developed by the working group on soil to plant transfer, International Union of Radioecology, France
111. Tebes-Stevens CL, Espinoza F, Valocchi AJ (2001) Evaluating the sensitivity of a subsurface multicomponent reactive transport model with respect to transport and reaction parameters. J Contam Hydrol 52:3–27
112. Steefel CI, DePaolo DJ, Lichter PT (2005) Reactive transport modeling: An essential tool and a new research approach for the earth sciences. Earth Planet Sc Lett 240:539–558
113. Davis JA, Yabusaki SB, Steefel CI, Zachara JM, Curtis GP, Redden GD, Criscenti LJ, Honeyman BD (2004) Assessing conceptual models for subsurface reactive transport of inorganic contaminants. EOS 85(44):449
114. Bossew P, Kirchner G (2004) Modelling the vertical distribution of radionuclides in soil. Part 1: The convection–dispersion equation revisited. J Environ Radioact 73:127–150
115. Kirchner G, Strebl F, Bossew P, Ehlken S, Gerzabek MH (2009) Vertical migration of radionuclides in undisturbed grassland soils. J Environ Radioact 100:716–720
116. Ivanov YA, Lewyckyj N, Levchuk SE, Prister BS, Firsakova SK, Arkhipov NP, Arkhipov AN, Kruglov SV, Alexakhin RM, Sandalls J, Askbrant S (1997) Migration of ^{137}Cs and ^{90}Sr from Chernobyl fallout in Ukrainian, Belarussian and Russian soils. J Environ Radioact 35(1):1–21
117. Pontedeiro EM, van Genuchten MT, Cotta RM, Simunek J (2010) The effects of preferential flow and soil texture on risk assessments of a NORM waste disposal site. J Hazard Mat 174:648–655
118. Gamerdinger AP, Kaplan DI (2000) Application of a continuous-flow centrifugation method for solute transport in disturbed, unsaturated sediments and illustration of mobile-immobile water. Water Resour Res 36(7):1747–1755
119. Seyfried MS, Rao PSC (1987) Solute transport in undisturbed columns of an aggregated tropical soil: Preferential flow effects. Soil Sci Soc Am J 51:1434–1444
120. Wellman DM, Gamerdinger AP, Kaplan DI, Serne RJ (2008) Effect of particle-scale heterogeneity on uranium(VI) transport in unsaturated porous media. Vadose Zone J 7(1):67–78
121. Zachara JM, Davis JA, McKinley J, Wellman D, Liu C, Qafoku N, and Yabusaki S (2005) Uranium geochemistry in vadose zone and aquifer sediments from the 300-area uranium plume, Pacific Northwest National Laboratory, Richland, WA, p 113

122. Pickens JF, Jackson RE, Inch KJ, Merritt WF (1981) Measurement of distribution coefficients using a radial injection dual-tracer test. Water Resour Res 17(3):529–544
123. Seaman JC, Bertsch PM, Miller WP (1995) Ionic tracer movement through highly weathered sediments. J Contam Hydrol 20:127–143
124. Seaman JC, Bertsch PM, Wilson M, Singer J, Majs F, Aburime SA (2007) Tracer migration in a radially divergent flow field: Longitudinal dispersivity and anionic tracer retardation. Vadose Zone J 6:373–386
125. Porro I, Newman ME, Dunnivant FM (2000) Comparison of batch and column methods for determining strontium distribution coefficients for unsaturated transport in basalt. Environ Sci Technol 34(9):1679–1686
126. Gamerdinger AP, Kaplan DI, Wellman DM, Serne RJ (2001) Two-region flow and decreased sorption of uranium (VI) during transport in Hanford groundwater and unsaturated sands. Water Resour Res 37(12):3155–3162
127. Maraqa MA, Wallace RB, Voice TC (1999) Effect of water saturation on retardation of ground-water contaminants. J Environ Eng 125(8):697–704
128. Phillippi JM, Loganathan VA, McIndoe MJ, Barnett MO, Clement TP, Roden EE (2007) Theoretical solid/solution ratio effects on adsorption and transport: Uranium(VI) and carbonate. Soil Sci Soc Am J 71:329–335
129. Reardon EJ (1981) Kd's - Can they be used to describe reversible ion sorption reaction in contaminant migration? Ground Water 19(3):279–286
130. Brusseau ML, Srivastava R (1999) Non-ideal transport of reactive solutes in heterogeneous porous media: 4. Analysis of the Cape Cod natural-gradient field experiment. Water Resour Res 35(4):1113–1125
131. Curtis GP, Davis JA, Naftz DL (2006) Simulation of reactive transport of uranium(VI) in groundwater with variable chemical conditions. Water Resour Res 42:WO4404 1–15
132. Ma R, Zheng C, Prommer H, Greskowiak J, Liu C, Zachara J, Rockhold M (2010) A field-scale reactive transport model for U (VI) migration influenced by coupled multirate mass transfer and surface complexation reactions. Water Resour Res 46:17p
133. Boggs MJ, Adams EE (1992) Field study in a heterogeneous aquifer 4. Investigation of adsorption and sampling bias. Water Resour Res 28(12):3325–3336
134. Seaman JC (1998) Retardation of fluorobenzoate tracers in highly weathered soil and groundwater systems. Soil Sci Soc Am J 62:354–361
135. McCarthy JF, Howard KM, McKay LD (2000) Effect of pH on sorption and transport of fluorobenzoic acid ground water tracers. J Environ Qual 29:1806–1813
136. Seaman JC, Bertsch PM, Korom SF, Miller WP (1996) Physicochemical controls on non-conservative anion migration in coarse-textured alluvial sediments. Ground Water 34(5): 778–783
137. Jacques D, Šimůnek J, Mallants D, and Van Genuchten MT (2005) Long term uranium in agricultural field soils following mineral P-fertilization, in the 10th International Conference on Environmental Remediation and Radioactive Waste Management Glascow, Scotland. p 6
138. Jacques D, Šimůnek J, Mallants D, Van Genuchten MT (2008) Modeling coupled hydrologic and chemical processes: Long-term uranium transport following phosphorus fertilization. Vadose Zone J 7(2):698–711
139. Smith JT, Elder DG (1999) A comparison of models for characterizing the distribution of radionuclides with depth in soils. J Europ Soil Sci 50(2):295–307
140. Cernik M, Federer P, Borkovec M, Sticher H (1994) Modeling of heavy metal transport in a contaminated soil. J Environ Qual 23:1239–1248
141. Bossew P, Gastberger M, Gohla H, Hofer P, Hubmer A (2004) Vertical distribution of radionuclides in soil of a grassland site in Chernobyl exclusion zone. J Environ Radioact 73:87–99
142. Bunzl K, Forster H, Kracke W, Schimmack W (1994) Residence times of fallout $^{239\ +\ 240}$Pu, ^{238}Pu, ^{241}Am and ^{137}Cs in the upper horizons of an undisturbed grassland soil. J Environ Radioact 22:11–27
143. Coughtrey PJ (1988) Models for radionuclide transport in soils. Soil Use Manag 73:87–99
144. Kirchner G (1998) Applicability of compartmental models for simulating the transport of radionuclides in soil. J Environ Radioact 38(3):339–352
145. Kent DB, Wilke JA, Davis JA (2007) Modeling the movement of a pH perturbation and its impact on adsorbed zinc and phosphate in a wastewater-contaminated aquifer. Water Resour Res 43(7):W07440
146. Stollenwerk KG (1998) Molybdate transport in a chemically complex aquifer: Field measurements compared with solute-transport model predictions. Water Resour Res 34(10):2727–2740
147. Parkhurst DL, Stollenwerk KG, Colman JA (2003) Reactive-transport simulation of phosphorus in the sewage plume at the Massachusetts military reservation, Cape Cod, Massachusetts. Air Force Center for Environmental Excellence, Northborough, MA, pp 1–33
148. Vaughan PJ, Shouse PJ, Goldberg S, Suarez DL, Ayars JE (2004) Boron transport within an agricultural field: Uniform flow versus mobile-immobile water model simulations. Soil Sci 169:401–412
149. Silva RJ, Nitsche H (1995) Actinide environmental chemistry. Radiochim Acta 70/71:377–396
150. Artinger R, Schuessler W, Scherbaum F, Schild D, Kim J (2002) ^{241}Am migration in a sandy aquifer studied by long-term column experiments. Environ Sci Technol 36(22): 4818–4823
151. Lu N, Kung KS, Mason CFV, Triay IR, Cotter CR, Pappas AJ, Pappas MEG (1998) Removal of plutonium-239 and americium-241 from Rocky Flats soil by leaching. Environ Sci Technol 32(3):370–374
152. Whicker W, Schultz V (1982) Radioecology: Nuclear energy and the environment. CRC Press, Boca Raton, FL

153. Sawhney BL (1966) Kinetics of cesium sorption by clay minerals. Soil Sci Soc Am J 30:565–569
154. Mahara Y (1993) Storage & migration of fallout strontium-90 and cesium-137 for over 40 years in the surface soil of Nagasaki. J Environ Qual 22:722–730
155. Zygmunt J, Chibowski S, Klimowicz Z (1998) The effect of sorption properties of soil minerals on the vertical migration rate of cesium in soil. J Radionchem Nucl Chem 231(1–2): 57–62
156. NCRP (2006) Cesium-137 in the environment: Radioecology and approaches to assessment and management. National Council on Raditation Protection and Measurement, Bethesda, MD, p 382
157. McKinley JP, Zeissler CJ, Zachara JM, Serne RJ, Lindstrom RM, Schaef HT, Orr RD (2001) Distribution and retention of ^{137}Cs in sediments at the Hanford Site, Washington. Environ Sci Technol 35(17):3433–3441
158. Comans RN, Haller M, Pretter PD (1991) Sorption of cesium on illite: Non-equilibrium behaviour and reversibility. Geochim Cosmochim Acta 55:433–440
159. Comans RNJ, Hockley DE (1992) Kinetics of cesium sorption on illite. Geochim Cosmochim Acta 56:1157–1164
160. Komarneni S (1978) Cesium sorption and desorption behavior of kaolinites. Soil Sci Soc Am J 42:531–532
161. deKoning A, Comans RNJ (2004) Reversibility of radiocaesium sorption on illite. Geochim Cosmochim Acta 68(13): 2815–2823
162. deKoning A, Konoplev AV, Comans RNJ (2007) Measuring the specific caesium sorption capacity of soils, sediments and clay minerals. Appl Geochem 22:219–29
163. Cremers A, Elsen A, DePreter P, Maes A (1988) Quantitative analysis of radiocaesium retention in soils. Nature 335: 247–250
164. McLean EO, Watson ME (1985) Soil measurement of plant-available potassium. In: Potassium in agriculture. American Society of Agronomy, Madison, WI, pp 277–308
165. Hsu CN, Chang KP (1994) Sorption and desorption behavior of cesium on soil components. Appl Radiat Isot 45(4): 433–437
166. Stauton S (1997) On the mechanisms which determine the fate of radiocaesium in soil. Analusis Mag 25:24–28
167. Pinder JE III, Garten CT Jr, Paine D (1980) Factors affecting radiocesium uptake by plants inhabiting a contaminated floodplain. Acta Ecologica 1(1):3–10
168. Hinton TG, Bell CM, Whicker FW, Philippi T (1999) Temporal changes and factors influencing ^{137}Cs concentration in vegetation colonizing an exposed lake bed over a three-year period. J Environ Rad 44:1–19
169. Seel JF, Whicker FW, Adriano DC (1995) Uptake of ^{137}Cs in vegetable crops grown on a contaminated lakebed. Health Phys 68:793–799
170. Hinton T, Knox A, Kaplan D, Serkiz S (2001) An in situ method for remediating ^{137}Cs-contaminated wetlands using naturally occurring minerals. J Radioanalytical Nucl Chem 249(1): 197–202
171. Pinder JE III, Hinton TG, Whicker FW, Smith JT (2009) Cesium accumulation by fish following acute input to lakes: A comparison of experimental and Chernobyl-impacted systems. J Environ Radioact 100:456–467
172. Hinton TG, Kaplan DI, Knox AS, Coughlan DP, Nascimento RV, Watson SI, Fletcher DE, Koo B (2006) Use of illitic clay for the in situ remediation of ^{137}Cs-contaminated water bodies: Field demonstration of reduced biological uptake. Environ Sci Technol 40(14):4500–4505
173. Leppert D (1990) Heavy metal sorption with clinoptilolite zeolite: alternatives for treating contaminated soil and water. Min Eng 42(6):604–608
174. Tsitsishvili GV, Andorikashvili TG, Kirov GN, Filizova LD (1992) Natural zeolites. Ellis Horwood, New York, 296
175. Zhang PC, Brady PV (eds) (2002) Geochemistry of soil radionuclides. Soil Science Societyof America, Madison, WI, p 252
176. Moran JE, Oktay S, Santchi PH, Schink DR (1999) Atmospheric dispersal of ^{129}I from nuclear fuel reprocessing facilities. Environ Sci Technol 33(15):2536–2542
177. Hou X, Aldahan A, Nielsen SP, Possnert G (2009) Time series of ^{129}I and ^{127}I speciation in precipitation from Denmark. Environ Sci Technol 43:6522–6528
178. Schwehr KA, Santschi PH, Kaplan DI, Yeager CM, Brinkmeyer R (2009) Organo-iodine formation in soils and aquifer sediments at ambient concentrations. Environ Sci Technol 43:7258–7264
179. Kaplan DI, Serne RJ, Parker KE, Kutnyakov IV (2000) Iodide sorption to subsurface sediments and illitic minerals. Environ Sci Technol 34:399–405
180. Hu Q, Zhao P, Moran JE, Seaman JC (2005) Sorption and transport of Iodine species in sediments from the Savannah River and Hanford Sites. J Contam Hydrol 78: 185–205
181. Li HB, Xu XR, Chen F (2009) Determination of iodine in seawater: Methods and applications. In: Comprehensive handbook of iodine: Nutritional, biochemical, pathological and therapeutic aspects. Elsevier Academic Press, New York, pp 2–13
182. Schwehr KA, Santschi PH (2003) Sensitive determination of iodine species, including organo-iodine, for freshwater and seawater samples using high performance liquid chromatography and spectrophotometric detection. Anal Chim Acta 482:59–71
183. Powell BA, Field RA, Kaplan DI, Coates JT, Serkiz SM (2004) Pu(V)O^{2+} adsorption and reduction by synthetic magnetite (Fe_3O_4). Environ Sci Technol 38:6016–6024
184. Powell BA, Field RA, Kaplan DI, Coates JT, Serkiz SM (2005) Pu (V)O^{2+} adsorption and reduction by synthetic hematite and goethite. Environ Sci Technol 39:2107–2114
185. Kim JI (1986) Chemical behaviour of transuranic elements in natural aquatic systems. Handbook on the Phys Chem Actinides 4:413–455
186. Kim JI, Kanellakopulos B (1989) Solubility product of Pu(VI) oxide and hydroxide. Radiochim Acta 48:145–150

187. Routson RC, Barney GS, Smith RM (1980) Hanford site sorption studies for the control of radioactive wastes: A review. Rockwell Hanford Operations, Richland, WA
188. Ames LL, McGarrah JE, Walker BA, Salter PF (1982) Sorption of uranium and cesium by Hanford basalts and associated secondary smectites. Chem Geol 35:205–225
189. Strenge DL, Peterson SR (1989) Chemical databases for the multimedia environmental pollutant assessment system, PNL-7145, Pacific Northwest National Laboratory Richland, WA
190. Serne RJ, LeGore VL (1996) Strontium-90 sorption-desorption properties and sediment characterization at the 100 -Area. Pacific Northwest National Laboratory, Richland, WA
191. Lefevre R, Sardin M, Schweich D (1993) Migration of strontium in clayey and calcareous sandy soil: Precipitation and ion exchange. J Contam Hydrol 13:215–229
192. Stevenson FJ, Fitch A (1986) Chemistry of complexation metal ions with soil solution organics. In: Huang PM, Schnitzer M (eds) Interactions of soil minerals with natural organics and microbes. Soil Science Society of America, Madison, WI
193. Wildung RE, McFadden KM, Garland TR (1979) Technetium sources and behavior in the environment. J Environ Qual 8(2):156–161
194. Kaplan DI, Roberts K, Shine G, Grogan K, Fjeld R, and Seaman J (2008) The range and distribution of technetium Kd values in the SRS subsurface environment, WSRC-STI-2008-00286 Aiken, SC. p 45
195. Peretyazhko T, Zachara JM, Heald SM, Jeon BH, Kukkadapu RK, Liu C, Moore D, Resch CT (2008) Heterogeneous reduction of Tc(VII) by Fe(II) at the solid-water interface. Geochim Cosmochim Acta 72:1521–1539
196. Peretyazhko T, Zachara JM, Heald SM, Kukkadapu RK, Liu C, Plymale AE, Resch CT (2008) Reduction of Tc(VII) by Fe(II) sorbed on Al (hydr)oxides. Environ Sci Technol 42:5499–5506
197. Wildung RE, Gorby YA, Krupka KM, Hess NJ, Li SW, Plymale AE, McKinley JP, Fredrickson JK (2000) Effect of electron donor and solution chemistry on products of dissimilatory reduction of technetium by shewanella putrefaciens. Appl Environ Microbiol 66(6):2451–2460
198. Murphy CE Jr (1993) Tritium transport and cycling in the environment. Health Phys 65(6):683–698
199. Okada S, Momoshima N (1993) Overview of tritium: Characteristics, sources and problems. Health Phys 65(6):595–609
200. Pointurier F, Baglan N, Alanic G, Chiappini R (2003) Determination of organically bound tritium background level in biological samples from a wide area in the south-west of France. J Environ Rad 68:171–189
201. van den Hoek J (1986) Tritium metabolism in animals. Radiat Prot Dosim 16:117–121
202. Moghissi AA, Bretthauer EW, Patzer RG (1987) Biological concentration of ^{3}H. Health Phys 53:385–388
203. Richmond CR, Langham WH, Trujillo TT (1962) Comparative metabolism of HTO by mammals. J Comp Physiol 59:45–53
204. Kelsey-Wall A, Seaman JC, Jagoe CH, Dallas CE (2006) Biological half-life and oxidative stress effects in mice with low-level, oral exposure to tritium. J Toxicology Environ Health Part A 69(3):201–213
205. Burns PC. and Finch R, eds. (1999) Uranium: Mineralogy, geochemistry, and the environment. Reviews in Mineralogy, vol 38. The Mineralogical Society of America: Washington, DC, p 679
206. Finn PA, Hoh JC, Wolf SF, Slater SA, Bates JK (1996) The release of uranium, technetium, and iodine from spent fuel under unsaturated conditions. Radiochim Acta 75:65–71
207. Vandenhove H, Hees MV, Wouters K, Wannijn J (2007) Can we predict uranium bioavailability based on soil parameters? Part 1: Effect of soil parameters on soil solution uranium concentration. Environ Pollut 145:587–595
208. Sandino A, Bruno J (1992) The solubility of $(UO_2)_3(PO_4)_2$-$4H_2O$ (s) and the formation of U(VI) phosphate complexes: their influence in uranium speciation in natural waters. Geochim Cosmochim Acta 56:4135–4145
209. Langmuir D (1978) Uranium solution-mineral equilibria at low temperatures with applications to sedimentary ore deposits. Geochim Cosmochim Acta 42:547–569
210. Artinger R, Rabung T, Kim JI, Sachs S, Schmeide K, Heise KH, Bernhard G, Nitsche H (2002) Humic colloid-borne migration of uranium in sand columns. J Contam Hydrol 58:1–12
211. Shi Z, Liu C, Zachara JM, Wang Z, Deng B (2009) Inhibition effect of secondary phosphate mineral precipitation on uranium release from contaminated sediments. Environ Sci Technol 43:8344–8349
212. Boult KA (1998) Towards an understanding of the sorption of U(VI) and Se(IV) on sodium bentonite. J Contam Hydrol 35:141–150
213. Sheppard SC (2001) Toxicants in the environment: Bringing radioecology and ecotoxicology together. In: Brechignac F, Howard BJ (eds) Radioactive pollutants: Impact on the environment. EDP Sciences, France, pp 63–74
214. Amonette JE, Holdren GR, Krupa KM, Lindenmeier CW (1994) Assessing the environmental availability of uranium in soils and sediments. Pacific Northwest Laboratory, Richland, WA, p 101
215. Fredrickson JK, Zachara JM, Kennedy DW, Duff MC, Gorby YA, Li SW, Krupka KM (2000) Reduction of U(VI) in goethite (α-FeOOH) suspensions by a dissimilatory metal-reducing bacterium. Geochim Cosmochim Acta 64(18):3085–3098
216. Lloyd JR, Lovley DR (2001) Microbial detoxification of metals and radionuclides. Curr Opin Biotechnol 12:248–253
217. Gu B, Wu W, Ginder-Vogel MA, Yan H, Fields MW, Zhou J, Fendorf S, Criddle CS, Jardine PM (2005) Bioreduction of uranium in a contaminated soil column. Environ Sci Technol 39(13):4841–4847
218. Cummings DE, Fendorf S, Singh N, Sani R, Peyton B, Magnuson T (2007) Reduction of Cr(VI) under acidic conditions by the facultative Fe(III)-reducing bacterium acidiphilium cryptum. Environ Sci Technol 41(1):146–152
219. Ginder-Vogel M, Criddle CS, Fendorf S (2006) Thermodynamic constraints on the oxidation of biogenic UO_2 by Fe (III) (hydr)oxides. Environ Sci Technol 40(11):3544–3550

220. Seaman JC, Hutchison J, Jackson BP, Vulava VM (2003) In situ treatment of metals in contaminated soils using phytate. J Environ Qual 32(1):153–161
221. Nash KL, Jensen MP, Schmidt MA (1998) Actinide immobilization in the subsurface environment by in-situ treatment with a hydrolytically unstable organophosphorus complexant: Uranyl uptake by calcium phytate. J Alloys Compounds 271–273:257–261
222. Nash KL, Jensen MP, Hines JJ, Friedrich SA, Redko M (1997) Phosphate mineralization of actinides by measured addition of precipitating anions. Chemistry Division-Argonne National Laboratory Argonne, IL, p 22
223. Nash KL, Jensen MP, Schmidt MA (1998) In-situ mineralization of actinides for groundwater cleanup: Laboratory demonstration with soil from the fernald environmental management project. In: Schulz WW, Lombardo NJ (eds) Science and technol. for disposal of radioactive tank waste. Plenum Press, New York, pp 507–518
224. Wellman DM, Pierce EM, Bacon DH, Oostrom M, Gunderson KM, Webb SM, Bovaird CC, Cordova EA, Clayton ET, Parker KE, Ermi RM, Baum SR, Vermeul VR, and Fruchter JS (2008) 300-Area treatability test: Laboratory development of polyphosphate remediation technology for in situ treatment of uranium contamination in the vadose zone and capillary fringe, PNNL, Richland, WA, p 110

Radionuclide Migration from Catchments, Modeling

LUIGI MONTE
ENEA, CR Casaccia, Rome, Italy

Article Outline

Glossary
Definition of the Subject and Its Importance
Introduction
Empirical Background
Principles for Developing Models of Radionuclide Migration from Catchments
Conclusions
Future Directions
Bibliography

Glossary

Catchment A delimited geographical area that collects water from rain and melting snow and ice flowing to a point of discharge into a water body.

Deposit Radionuclide inventory in a catchment, often expressed as radionuclide per square meter ($\mathrm{Bq\,m^{-2}}$).

Radionuclide deposition The process of migration to exposed surfaces of a radionuclide dispersed in the atmosphere.

Radionuclide washout The process of migration of a radionuclide transported by runoff waters from the catchment to a water body.

Runoff The portion of precipitation on a catchment that is discharged into a water body.

Transfer function (of radionuclide deposited over a catchment) The amount of radionuclide flowing per unit time from upstream drainage basin to a water body following a single-pulse deposition of radioactive substance.

Definition of the Subject and Its Importance

A catchment, or drainage basin, is a delimited geographic area that collects water from rain and melting snow and ice flowing to a point of discharge into a water body. Catchments are complex ecosystems of significant economic, social, and environmental value. They play a major role in determining the water quality of streams, rivers, lakes. The portion of waters flowing through the catchment that reaches the point of discharge, the so-called runoff, transports dissolved and particulate substances, such as minerals, nutrients, eroded soil particles, that significantly influence the chemical and the biological characteristics of the water body that receives the runoff.

Catchments comprise terrestrial and aquatic ecosystems including forests, grasslands, ponds, etc., whose behavior is influenced by a great many natural and nonnatural factors that depend on geographic, geological, and climatic conditions and on the impact on the environment of the human activities.

The flow and the balance of the water in a catchment are controlled by many complex processes such as the interception of precipitation by the above ground vegetation, evapotranspiration (the conversion to the vapor state of waters on surfaces, soils, vegetation), storage in catchment depressions, snow and ice accumulation, infiltration through the soil and the rocks, overland flow, interflow, and the groundwater flow through the saturated and the unsaturated zones of the watershed.

Contaminants deposited over the catchment surface can be transported to the water ecosystems contributing to persistent levels of pollution in rivers, lakes, coastal areas, and seas. The complex processes of radionuclide migration through a drainage area depend on the characteristics of the radioactive substances and of the catchment. More specifically, these processes are chiefly controlled by the hydrology of the drainage basin and by the mechanisms of interaction of radionuclides with soils and bedrocks.

The radioactive pollution of a catchment can be caused by events occurring at different scales. However, the contamination of large geographical areas following the introduction of radionuclides into the atmosphere is of particular significance in view of the possible radioecological impact on water bodies and the relevant consequences for the human health. On the medium and long term, radionuclides migrating from catchments can be important sources of contamination for the water bodies and, consequently, of radiation doses to man.

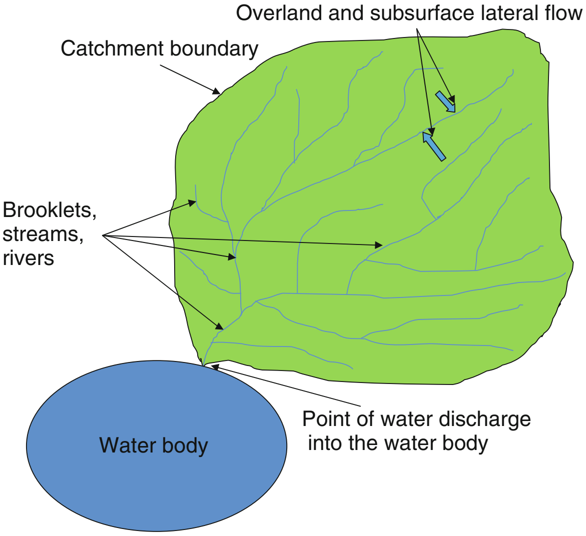
In particular, the global fallout of radionuclides from the nuclear weapon tests in atmosphere carried out in the early 1960s and severe cases of environmental pollution like the Chernobyl accident caused the radioactive contamination of large catchments.

Modeling the transfer of radionuclides from drainage areas to water bodies is essential to evaluate the consequences of accidental releases of radioactive substance into the environment and to develop environmental management plans in view of the social, economic, and ecological values of freshwater ecosystems.

A variety of models for predicting the migration of toxic substances from catchments have been developed. The main aim of these models is to determine, following the introduction of contaminants into the catchment, the amount of radionuclides migrating, per unit time, from the drainage area to the freshwater body or to the marine environment. This is a major challenge for modelers owing to the complicated web of processes that control the migration of toxic substances through the complex freshwater environment.

Introduction

The complex hydrological processes (Fig. 1) controlling the runoff from a drainage area depend on the topography, the geology, the environmental conditions, etc., of the catchment.

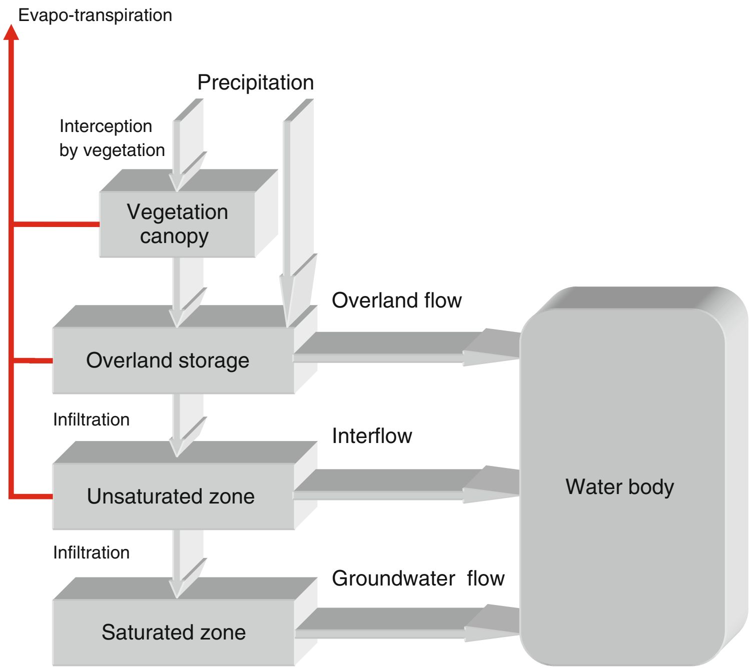

Overland and subsurface lateral flow
Catchment boundary
Brooklets, streams, rivers
Water body
Point of water discharge into the water body

Radionuclide Migration from Catchments, Modeling. Figure 1
Simplified scheme of a catchment and of the main water fluxes

For the purpose of the present entry, it is sufficient to pay attention to the schematic representation of the main water fluxes depicted in Fig. 2. Precipitation is intercepted by the vegetation cover and by the soil surface, then infiltrates from the water unsaturated zone (strata of rocks, sand, gravel, etc., where available spaces are not completely filled with water) and percolates to the saturated zone. A portion of the water is returned to the atmosphere by evaporation and transpiration, whereas the remaining part flows to water bodies.

Runoff waters that reach the point of discharge carry dissolved substances and particulate materials of natural origin, but can also transport contaminants that have been released into the environment.

The evaluation of the transfer of contaminants from catchments to aquatic ecosystems is of paramount importance for the proper assessment of the aftermath of events of accidental introduction of pollutants into the biosphere. In particular, the pollution of large geographical areas following the release of pollutants into the atmosphere can be an event with severe consequences for aquatic ecosystems.



Radionuclide Migration from Catchments, Modeling. Figure 2
Main water fluxes that control the migration of toxic substances through a catchment

A contaminant discharged into the atmosphere is deposited on the ground surface through processes of particle settling (dry deposition), through the direct adsorption onto exposed surfaces of pollutants in the gaseous phase, and, chiefly, as wet deposition caused by precipitation scavenging. A further mechanism of contaminant migration to the ground is the so-called cloud deposition that is caused by the deposition of contaminated droplets in low clouds and mist on exposed surfaces. Generally, this last mechanism, although dissimilar from precipitation scavenging, is also classified as wet deposition. The deposited contaminant is partially adsorbed onto the exposed surface in the catchment (vegetation cover, rocks, soils, etc.) and is transported by runoff waters to the point of discharge in dissolved form or attached to eroded particles (radionuclide washout).

Basically, the migration of a contaminant within the catchment is governed by water flow, sediment erosion, and the complex physical and chemical processes of contaminant interaction with the exposed surfaces in the catchment (in particular, with the soil particles). These processes depend on the characteristics of the catchments and of the pollutant.

Radionuclides transported by runoff waters are often transformed into different chemical/physical forms ("speciation") following the interaction with different kinds of rocks and soil particles in a complex geochemical environment. However, it is commonly assumed that these forms can be grouped in two broad categories, "particulate" and "dissolved," according to a simplified and conventional approach based on an operational definition of "dissolved substances" – the material passing a 0.45 µm membrane filter [1] (correspondently, the "particulate form" is the suspended material that can be collected by such filters). Schematically, radionuclides are transported throughout the catchment into the water body by: (a) the overland flow, (b) the interflow (the water flowing through the upper part of the unsaturated zone), and (c) the groundwater flow. As it will be described in the following sections, the above mentioned flows contribute in different measure to the contaminant washout, depending on the chemical properties of the radionuclide that control the interaction of the pollutant with the particulate matter and the rocks in the catchment.

The amount of radionuclide migrating from the catchment is a significant component of the mass-balance equation for predicting the time behavior of the radioactive contamination of a water body. The mass conservation law requires that:

$$\frac{dT}{dt} = DR + IR + CR - OR - RDR \qquad (1)$$

where

T is the total amount of radionuclide in the water body (Bq)
DR is the rate of deposition of radionuclide over the water body surface
IR is the rate of direct introduction of radionuclide into the water body (from sources other than the water body catchment)
CR is the rate of radionuclide migrating from the catchment to the water body
OR is the rate of radionuclide outflow from the water body (for instance, through the outlet of a lake)
RDR is the radioactive decay rate of the contaminant in the water body
t is time

All the terms in the right-hand side of the previous equation are expressed in $\mathrm{Bq\,s^{-1}}$.

In particular, the contribution of contaminant from the catchment of an aquatic system characterized by a relatively short mean water retention time, such as a river, can be one of the most important components of the contaminant mass balance in the water body.

The present contribution focuses on the description and the analysis of the most common methodological techniques to develop models for predicting the migration of radionuclides from catchments to water bodies. Before discussing the modeling approaches, the empirical background and the main methods adopted to analyze the available experimental data will be reviewed. This can help in understanding the motivation of the principles underlying the different classes of models described.

Empirical Background

Experimental Results Following the Environmental Contamination Caused by the Nuclear Weapon Explosions in the Atmosphere

Since the beginning of the 1960s, a great deal of research has been carried out to investigate the transfer of radionuclides from catchments to water bodies on the basis of available experimental data.

The worldwide fallout of radioactive substances following the peacetime testing of nuclear weapons was a source of copious experimental data for studying the process of migration of radionuclides through the environment. The radioactive debris released by the nuclear explosions reached the troposphere and the stratosphere, were widely dispersed through the atmosphere and returned to the earth as "global fallout" [2, 3].

Research on the migration of radionuclides from catchments was initiated as soon as the fallout from nuclear weapon tests in atmosphere caused the contamination of large geographical areas [4, 5]. These studies were aimed at establishing possible relationships between the radionuclide concentration in runoff water at the point of water discharge and the contaminant deposition over the catchment.

Yamagata et al. [5] tried to evaluate the washout of ^{137}Cs and ^{90}Sr deposited over the catchments of rivers in Japan by fitting empirical data of radionuclide concentration *C* ($\mathrm{Bq\,m^{-3}}$) in river water to the expression:

$$C = C_d + p_r \cdot C_r \qquad (2)$$

where C_r ($\mathrm{Bq\,m^{-3}}$) is the concentration of radionuclide in the rain water, p_r (dimensionless) is a proportionality factor ($p_r \cdot C_r$ corresponds to the process of instantaneous transfer of deposited contaminant to the runoff water) and C_d ($\mathrm{Bq\,m^{-3}}$) is the radionuclide concentration in the river water associated with the cumulative deposition stored in the catchment. Equation 2 was based on a simple conceptual scheme of the mechanisms of transfer from the whole catchment to the water body of deposited radionuclides. The results of the regression analyses suggested that, approximately, 1.3% of deposited ^{137}Cs and 7.2% of deposited ^{90}Sr were quickly removed from the catchment.

Furthermore, let *D* denote the cumulative radionuclide deposit on the ground ($\mathrm{Bq\,m^{-2}}$), *F* the radionuclide runoff from the catchment ($\mathrm{Bq\,s^{-1}}$), λ_w ($\mathrm{s^{-1}}$) the removal rate of the cumulative ground deposit and *A* ($\mathrm{m^2}$) the surface area of the catchment. Then it is possible to write:

$$F = \lambda_w \cdot D \cdot A \qquad (3)$$

From the application of Eq. 3 to the available empirical data, Yamagata et al. [5] estimated that the removal rates λ_w of ^{137}Cs and ^{90}Sr from catchments were of the order of 1.9×10^{-11} s^{-1} and 9.8×10^{-11} s^{-1}, respectively, concluding that those values were significantly lower than the radioactive decay constants.

It is worthwhile to note that the deposition caused by the nuclear weapon tests in the atmosphere was a continuous process. It was impossible to isolate the input of radionuclide into the environment from each single explosion. Moreover, the dynamics of deposition over the ground surface of radionuclides from an atmospheric nuclear weapon test showed a complex time behavior as a significant amount of contaminant was introduced into the stratosphere and returned slowly to the troposphere and the earth [2].

The main problems with the above mentioned evaluations were the difficulty of determining the transfer rates from data of environmental contamination caused by the dynamics of continuous deposition and, chiefly, the high uncertainty levels of the estimates of radionuclide deposition on the whole catchment and of the contaminant runoff.

Experimental Results Following the Chernobyl Accident

Following the accident on 26 April 1986 at the Chernobyl Nuclear Power Plant located in Ukraine, radioactive substances were dispersed into the environment. The Chernobyl accident represented a line of demarcation between research carried out following the environmental contamination due to the nuclear weapon tests in the atmosphere and the new results from a single contamination event. The releases of contaminants occurred over a few weeks period and the radionuclides spread across Europe. Vast regions were affected by the consequences of the accident and whole large river catchments were contaminated [6]. The accident can be classified as a pulse-type event of radionuclide release, at least in relation to those environmental migration processes that occur over periods of time of the order of months or years such as the migration on the medium and long term of radionuclides from a catchment.

Following the accident, a variety of studies focused on the evaluation of the quantitative behavior of radionuclide in catchments [7–9]. Some results from the analysis of contamination data collected by various European laboratories [10–17] are analyzed and discussed below.

The main assumption of the above mentioned studies was that, following a pulse of radionuclide deposition D (Bq m^{-2}) occurred at time $t=0$, the flux of radionuclide, $F(t)$ (Bq s^{-1}), transported by the river water at instant t can be expressed as D multiplied by a “radionuclide transfer function per unit deposition” $TF(t)$ (m^2 s^{-1})

$$F(t) = D \cdot TF(t) \tag{4}$$

Generally, function TF was fitted to the sum of time-dependent exponential components $e^{-(\lambda+\lambda_i)t}$ which account for environmental effects and other first-order processes, such as the radioactive decay, that influence the time behavior of the pollutant concentration in the water flowing through the catchment:

$$TF(t) = \sum_i A_i^* \, e^{-(\lambda+\lambda_i)t} \tag{5}$$

A_i^* are multiplicative factors, λ_i are empirical parameters controlling the decay of radionuclide concentration in water due to environmental effects (s^{-1}) and λ is the radioactive decay constant (s^{-1}).

We can assume that, in given hydrological conditions (water fluxes, amount and characteristics of suspended matter in water, environmental conditions due to seasonal effects, such as the fraction of catchment area covered by snow, ice, etc.), the initial concentration $C(0)$ of radionuclide in water is proportional to the deposition pulse D:

$$C(0) = \varepsilon \, D \tag{6}$$

where ε is a coefficient (m^{-1}) that depends on the radionuclide, on the catchment characteristics, and on the above mentioned conditions. In principle, coefficients A_i^* may depend on the water fluxes $\Phi(t)$ and the relationships between A_i^* and $\Phi(t)$ can be nonlinear. It is useful to “normalize” the water flux and the radionuclide migration rate to the initial conditions. By such a “normalization,” it is possible to fit the empirical data of contaminant flux to the function $(F(t)/F(0))$ that characterizes the time behavior of the radionuclide transfer from the catchment:

$$\frac{F(t)}{F(0)} = \sum_i A_i \left(\frac{\Phi(t)}{\Phi(0)}\right)^{\alpha_i} e^{-(\lambda_i+\lambda)t} \tag{7}$$

α_i are empirical exponents (dimensionless) that give reason for the above mentioned possible nonlinearity of $F(t)$ as a function of the water flux from the catchment. Coefficients A_i (dimensionless) are the relative weights of the exponential components in the transfer function.

The use of the ratios $(F(t)/F(0))$ and $(\Phi(t)/\Phi(0))$ assures that coefficients A_i are dimensionless and that $\sum_i A_i = 1$, as it can be easily demonstrated by putting t=0 in Eq. 7.

The flux of radionuclide at instant t can be simply expressed in terms of D and ε. Indeed, in view of

$$C(t) = \frac{F(t)}{\Phi(t)} \tag{8}$$

and of Eq. 6, we can write for t=0:

$$F(0) = \varepsilon D \Phi(0) \tag{9}$$

Consequently, by accounting for Eq. 9, Eq. 7 is transformed as follows:

$$F(t) = \varepsilon D \sum_i [\Phi(t)]^{\alpha_i} \beta_i A_i e^{-(\lambda_i+\lambda)t} \tag{10}$$

or, in view of Eq. 8,

$$C(t) = \varepsilon D \sum_i [\Phi(t)]^{\alpha_i - 1} \beta_i A_i e^{-(\lambda_i+\lambda)t} \tag{11}$$

where

$$\beta_i = [\Phi(0)]^{1-\alpha_i} \tag{12}$$

are "normalization" coefficients

The main advantage of formula (7) is that Eq. 10 can be fitted to the experimental data of radionuclide fluxes by using the empirical dimensionless estimates of variables $(F(t)/F(0))$ and $(\Phi(t)/\Phi(0))$ in order to determine the values of parameters ε, λ_i, A_i, and α_i.

If the effects of the seasonal variations of the water flow on the radionuclide concentration in water are not accounted for (for instance, if the data analysis is based on yearly average of the water fluxes), we can put α_i=1. With such an assumption, from Eqs. 11 and 12, we obtain

$$C(t) = \varepsilon D \sum_i A_i e^{-(\lambda+\lambda_i)t} \tag{13}$$

As Eq. 13 shows, the yearly average concentration in water comprises two terms:

– A multiplicative "scaling factor" ε
– The sum:

$$\sum_i A_i e^{-(\lambda+\lambda_i)t} \tag{14}$$

Expression (14) characterizes the "shape" of the curve (i.e., the time behavior of the migration process).

The scaling factor ε, the ratio between the initial concentration of radionuclide in water and the total deposition, determines the "position" of the curve in the graph "radionuclide concentration in water versus time" [18]. From the available experimental data collected over the course of 5–7 years following the Chernobyl accident, it was possible to identify short-term and long-term components for periods of few months and of several years, respectively.

The above described notion of transfer function can be applied to the dissolved, the particulate, and the total (particulate+dissolved) concentration of radionuclide in water. A review of available empirical evaluations of the effective decay constants, of the coefficient A_2 and of the exponent α_2 are reported in Table 1.

Examples of applications of Eq. 11 are shown in Figs. 3 and 4. The concentrations of ^{137}Cs and ^{90}Sr in the River Prypiat (Ukraine) are reported as functions of time and are compared with the values calculated by Eq. 11.

The values of α_2 are higher than 1 for strontium, a radionuclide that is chiefly in dissolved form. This suggests that the concentration in water of such a radionuclide is positively correlated with the water flux. A possible explanation of this effect is the different contributions of dissolved radionuclide from soil layers at different depth. Following deposition of a radionuclide, the upper part of soil is heavily contaminated and significantly contributes to the contamination of the surface runoff waters. When the flux of water in the catchment increases due to the effects of heavy precipitation or snow melting, the water saturation of soils in the catchment implies that a larger fraction of water flows as surface runoff and interacts with the more contaminated upper soil layers. There is no similar effect for dissolved cesium, (α_2 ranges from 0.53 to 1.08 and, in most cases, the correlation coefficient

Radionuclide Migration from Catchments, Modeling. Table 1 Measured values of some parameters in the transfer function

River	Radionuclide	Period of collection of fitted data (following the accident)	A_2 (dimensionless)	α_2 (dimensionless)	Standard deviation of α_2	$\lambda_1+\lambda$ (s^{-1})	Standard deviation of $\lambda_1+\lambda$	$\lambda_2+\lambda$ (s^{-1})	Standard deviation of $\lambda_2+\lambda$	Reference
Po[a]	^{137}Cs	Few months				2.3×10^{-7}	5.5×10^{-8}			[9][c]
Rhine[a]	137Cs	< 2 years	0.052	0.53	0.3	6.5×10^{-7}	1.3×10^{-7}	2.7×10^{-8}	0.6×10^{-8}	[9]
Prypiat[a]	^{137}Cs	≈5 years	0.035	1.08	0.06	5.2×10^{-7}	6.5×10^{-7}	1.8×10^{-8}	0.7×10^{-9}	[9]
Dnieper[a]	^{137}Cs	≈5 years	0.028	0.86	0.06	8.8×10^{-7}	1.1×10^{-7}	1.1×10^{-8}	0.7×10^{-9}	[9]
Teterev[a]	^{137}Cs	≈5 years		0.96	0.15			8.2×10^{-9}	$2. \times 10^{-9}$	[9]
Uzh[a]	^{137}Cs	≈5 years		1.02	0.1			1.5×10^{-8}	1.8×10^{-9}	[9]
Inlets of Devoke Water[b]	^{137}Cs	few years		1.0–1.3				1.2×10^{-8}		[7]
Inlets of lakes	^{137}Cs	≈6 years				$.6 \times 10^{-7}$– 1.5×10^{-7}		$7. \times 10^{-9}$– $2. \times 10^{-8}$		[13]
Hillesjön and Salgsjön[b] Garonne, Meuse, Moselle, Rhône, Seine	^{137}Cs	≈18 years						3.6×10^{-9} $-8. \times 10^{-9}$		[14][d]
Kemijoki, Kymijoki Tornionjoki, Kokemaenjoki, Oulujoki[b]	^{137}Cs	≈15 years						1.2×10^{-8} -4.9×10^{-8}		[15]

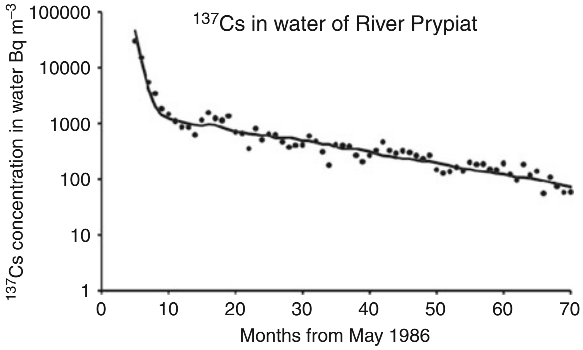
Po[a]	^{131}I	Few months				1.1×10^{-6}	6.5×10^{-8}			[9][c]
Po[a]	^{103}Ru	Few months				4.7×10^{-7}	4.0×10^{-8}			[9][c]
Prypiat	^{90}Sr	≈5 years	0.048	1.55	0.08	9.0×10^{-7}	1.1×10^{-7}	4.9×10^{-9}	0.9×10^{-9}	[9]
Dnieper	^{90}Sr	≈5 years	0.166	1.4	0.08	5.2×10^{-7}	1.5×10^{-7}	5.5×10^{-9}	0.9×10^{-9}	[9]
Teterev	^{90}Sr	≈5 years		1.12	0.14			3.6×10^{-9}	2.1×10^{-9}	[9]
Uzh	^{90}Sr	≈5 years		1.31	0.09			5.9×10^{-9}	1.8×10^{-9}	[9]
Irpen	^{90}Sr	≈15 years						1.6×10^{-9}	–	[16]
Ilya	^{90}Sr	≈15 years						2.7×10^{-9}	–	[16]
Sakhan	^{90}Sr	≈15 years						3.8×10^{-9}	–	[16]
Glinitsa	^{90}Sr	≈15 years						1.9×10^{-9}	––	[16]

[a] Dissolved radionuclide

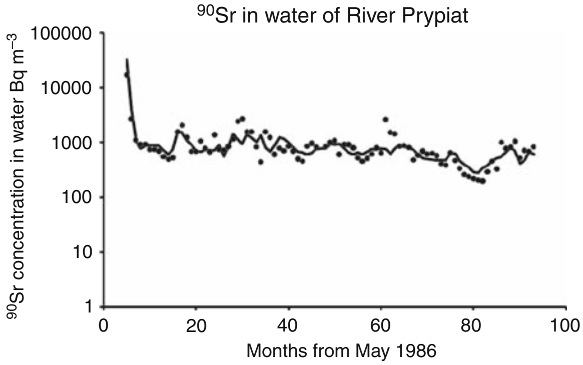
[b] Total ^{137}Cs (particulate+dissolved)

[c] Data fitted to a single exponential function

[d] Assesses from contamination decline in sediment and biota



Radionuclide Migration from Catchments, Modeling. Figure 3
Empirical values of ^{137}Cs concentration in water of River Prypiat (dissolved form) compared with the results of Eq. 11



Radionuclide Migration from Catchments, Modeling. Figure 4
Empirical values of ^{90}Sr concentration in water of River Prypiat compared with the results of Eq. 11

between ^{137}Cs in dissolved form and the water flux is <0.1). This is due to the fast fixation of these radionuclides to soil particles and to the dilution in the river waters that, generally, show lower levels of contamination. Indeed, a significant amount of the radionuclide is removed from the water column as a consequence of the intense interaction with settling suspended matter and bottom sediments.

Table 2 shows the values of some measured parameters of the transfer function, TF, for particulate cesium. It is worthwhile to note that the values of α_2 are higher than 1 probably as a consequence of the significant amounts of suspended matter in river water during periods of high water fluxes.

As Tables 1 and 2 show, several experimental evaluations of the effective decay rates and of the values of the exponential components of the TF were obtained. However, comparatively few experimental assessments are available for the transfer coefficient ε. Among the parameters in the transfer function, ε is the one that most significantly influences the uncertainty of the predictions. This parameter depends on many factors of environmental significance. Consequently, its values vary within a wide range. It is extremely important to evaluate ε for the specific environmental conditions of the catchment in order to perform reliable assessments of the radionuclide migrating from the drainage area. Smith et al. [18] analyzed the correlations between ε and several environmental characteristics of the catchments of 25 rivers in Europe and Asia. The analysis showed a significant dependence of ε on the percentage of "inland water" in the catchment. The "inland water" percentage (I_{WP}) values were obtained from the Advanced Very High Resolution Radiometer (AVHRR) database. As high values of reflectance from water surface correspond to very wet land areas, the "inland water" percentage provides a good estimate of the area of a catchment that can be more actively involved in the contaminant migration owing to possible higher levels of surface runoff. A linear relation between ε and I_{WP} was suggested:

$$\varepsilon = a \cdot I_{WP} + b \tag{15}$$

The estimated values of the multiplicative coefficient "a" were 0.011 and 0.14 m^{-1} for ^{137}Cs and ^{90}Sr, respectively. The values of "b" were 0.063 m^{-1} (^{137}Cs) and 0.55 m^{-1} (^{90}Sr). The values for strontium are higher than the corresponding values for cesium as the former radionuclide is more mobile than the latter in the aquatic environment.

Principles for Developing Models of Radionuclide Migration from Catchments

The migration of radioactive substances from drainage areas is influenced by complex processes of hydraulic, geochemical, sedimentological, ecological, and anthropogenic nature. Broadly speaking, a model for

Radionuclide Migration from Catchments, Modeling. Table 2 Measured values of some parameters in the transfer function from catchments (particulate cesium) [17]

River	α_2	95% up of α_2	95% down of α_2	$\lambda_2+\lambda$ (s^{-1})	95% up of $\lambda_2+\lambda$ (s^{-1})	95% down of $\lambda_2+\lambda$ (s^{-1})
Danube	2.44	1.90	2.98	1.4×10^{-8}	2.2×10^{-8}	6.7×10^{-9}
Uzh	1.02	0.65	1.39	1.1×10^{-8}	1.8×10^{-8}	4.0×10^{-9}
Teterev	1.34	0.97	1.77	1.2×10^{-8}	1.8×10^{-8}	6.0×10^{-9}
Prypiat	1.52	1.34	1.7	1.4×10^{-8}	1.6×10^{-8}	1.3×10^{-8}
Dnieper	1.24	1	1.37	1.2×10^{-8}	1.4×10^{-8}	1.1×10^{-8}
Desna	1.11	0.83	1.39	8.9×10^{-8}	1.3×10^{-8}	4.7×10^{-9}
Rhine	1.12	0.27	1.97	1.7×10^{-8}	2.4×10^{-8}	1.0×10^{-8}
Geometric mean	1.34			1.2×10^{-8}		

predicting the migration of radionuclides from catchments should account for [19]:

1. Hydrological processes
   (a) Water fluxes
   (b) Sediment erosion, dispersion, and deposition
2. Physicochemical processes of interaction of radionuclides with the exposed surfaces in the catchment and, in particular, with soils and eroded particles and the consequent partition of the contaminant in dissolved and particulate phases
3. The migration of dissolved and particulate contaminants transported by the waters flowing through the catchment

A great deal of work has been done in past decades to investigate the above processes.

The hydrologic cycle and the surface water runoff and balance have been thoroughly studied and many models have been developed. The evaluations of water runoff, of erosion rates and of the fluxes of water within the various vertical horizons of the catchment can be performed, at least in principle, by using equations that are copiously described in the scientific literature [20, 21].

The physicochemical processes of the interaction of radionuclides dissolved in water with soil and bedrock have been the subject of many studies in past decades [22, 23]. An extensive review of similar investigations, whose results chiefly refer to nonradioactive substances but that can be also applied, at least conceptually, to radionuclides, can be found in the specialized literature [24].

Finally, the evaluation of the migration of radionuclides transported, in dissolved and particulate forms, by waters flowing through the catchment is based on the application of the contaminant mass balance equation and on the assessment of the complex processes of radionuclide dispersion such as diffusion and transport [25].

The overwhelming complexity of the previously mentioned processes requires the application of suitable strategies to develop models for practical use [26–28].

In principle, the models for predicting the behavior of complex environmental processes can be classified as reductionistic or aggregated.

As all the above mentioned classes of hydrological, physicochemical, and migration processes have been copiously studied, it seems, at first sight, that the appropriate linking of the corresponding sub-models makes it possible to develop a comprehensive general model of the behavior of radionuclides migrating from catchments to water bodies. This is the well-known principle that governs the so-called reductionistic approach for modeling natural phenomena.

Reductionistic models are aimed at predicting the behavior of complex environmental systems by accounting for as many processes as reasonably possible. The term reductionism emphasizes the fact that the functioning of the considered system is modeled according to primary laws from fundamental disciplines such as physics and chemistry. [29].

A reductionistic model, in principle, accounts for a great deal of processes and, consequently, makes use of a copious number of parameters and variables in the equations controlling these processes. For instance, one of these models [30] predicts the migration of radiocaesium from lake catchments by accounting for the following processes and the relevant parameters: interception of rain and radionuclides by vegetation; infiltration through soil layers; effects of catchment slope, of soil density, porosity, and soil particle size; effects of washout from vegetation; etc. The model includes sub-models to determine the overload flow, the water depth of channeled flow, the water flow velocity, the overland flow of particulate matter, etc.

An alternative method consists in developing models based on the aggregation of processes and parameters. The aggregation occurs not only at level of processes but also by spatial integration over entire subregions of the catchment. It is assumed that the behavior of a complex system can be predicted, with acceptable approximations (at least in relation to those target variables that are deemed of interest in view of the specific problem), by means of more or less simple models that make use of a limited set of information, data, and parameters to describe the dynamics of the contaminant in the environment.

In the following sections, the main principles of the above mentioned modeling strategies will be described and discussed.

Principles Underlying Reductionistic Models

Several reductionistic models for predicting the fluxes of radionuclides from complex catchments to water bodies have been developed. These models simulate the dynamics of waters flowing to the water body and the resulting radionuclide migration. In principle, they are based on complicated hydrological equations, accounting for precipitation, water infiltration, interception of rain by the vegetation canopy, evapotranspiration, etc., and assessing the consequent dispersion of radionuclides through the different catchment components [31–33].

Reductionistic models make use of many parameters that are frequently assessed by specific sub-models and account for the spatial variability of the catchment hydrologic and geophysical characteristics. In several cases, appropriate hypotheses make it possible to simplify the approaches for evaluating the complex dynamics of the many involved processes.

The application of a reductionistic model requires the knowledge of the catchment characteristics and of the model parameters at levels of spatial resolution appropriate for modeling the processes with the required precision. It is beyond the scope of this entry to describe in detail such models. The main aim of the present section is to review the most important principles underpinning the reductionistic modeling approach.

It is worthwhile to remind the reader of the meaning of certain terms as to their *particular usage* in this entry. "Transport" is the ordered movement of the contaminant transported by the water current, while "dispersion" denotes the disordered movement of the pollutant caused by chaotic processes such as the molecular diffusion and the turbulent motion of water. The expression "radionuclide migration" defines the overall process of contaminant movement caused by both transport and dispersion processes.

A reductionistic model accounts for the dispersion and the transport of radionuclides through the saturated and the unsaturated zones of the aquifer (Fig. 2).

The dispersion implies the movement of a pollutant from regions of space of high contaminant concentration to regions of low concentration. The radionuclide flux, $\mathbf{F}$ ($\mathrm{Bq\,m^{-2}\,s^{-1}}$) (bold characters denote vectors or tensors), is related to the concentration gradient according to the Fick's first law:

$$\mathbf{F} = -\mathbf{K}\ \mathbf{grad}C \tag{16}$$

where **grad** is the gradient of C ($\mathbf{grad}C = \mathbf{i}\frac{\partial}{\partial x}C + \mathbf{j}\frac{\partial}{\partial y}C + \mathbf{k}\frac{\partial}{\partial z}C$), C ($\mathrm{Bq\,m^{-3}}$) is the dissolved radionuclide concentration in water and **i**, **j**, and **k** are the unit vectors along the coordinate axes x, y, and z. **K** is a 3×3 component symmetric tensor (the diffusion tensor).

Tensor **K** in Eq. 16 highlights that, in principle, diffusion is a non-isotropic process occurring in the three dimensions.

The radionuclide flux due to the ordered motion caused by the water current is related to the concentration C of radionuclide and to the water velocity $\mathbf{V}$ (ms^{-1}) by the following equation:

$$\mathbf{F} = \mathbf{V}\ C \tag{17}$$

From the Mass Conservation Law and Eqs. 16 and 17 that control the contaminant flux, it is possible to derive the equation governing the behavior of the concentration of a radionuclide in the water flowing through the catchment.

Let θ (m^3 m^{-3}) denote the volumetric content of water, at point of coordinate (x, y, z), in the substrate (for instance, the sand, the gravel, etc.) that adsorbs the contaminant. The value of θ depends on the volume of water filling the pores of the substrate components. It is lower than 1 m^3 m^{-3} and reaches this maximum value in surface water where the volume of particulate matter per unit volume is negligible. The total flux $\mathbf{F}$ (Bq m^{-2}s^{-1}) of radionuclide in dissolved form at point x is:

$$\mathbf{F} = (-\mathbf{K} \times\ \mathbf{grad}C + \mathbf{v}\ C)\theta \tag{18}$$

The application of the contaminant mass balance equation over a finite region of space and the following transformation of the surface integral of flux $\mathbf{F}$ in a volume integral yield, in view of Eq. 18:

$$\frac{\partial}{\partial t}(\theta C + \delta C_s) + \lambda(\theta C + \delta C_s) + div\mathbf{F} = 0 \tag{19}$$

where C_s (Bqkg^{-1}) is the concentration of radionuclide attached to the firm substrate, δ (kgm^{-3}) is the mass of substrate per cubic meter and λ is the radioactive decay constant.

To determine C and C_s from Eq. 19 requires an appropriate sub-model for predicting the dynamics of the interaction of radionuclide in dissolved form with particulate matter.

The most common and simple approach to model this interaction with substrate matter is based on the hypothesis of a reversible rapid equilibrium between the dissolved and the adsorbed phases. More specifically, if the substrate is composed of particulate matter like sand or clay, the ratio between the radionuclide concentration in particulate form (Bq per unit mass of particulate) and the radionuclide in dissolved form (Bq per unit volume of water) is the so-called distribution or partition coefficient k_d (m^{3}kg^{-1}) [34]:

$$\frac{C_s}{C} = k_d \tag{20}$$

The distribution coefficient may be used as a first approach to evaluate the behavior of pollutants like radionuclides and heavy metals in the system water-soil when the ratio of the pollutant concentration in soil (particulate phase) and the pollutant concentration in interstitial water (dissolved phase) quickly reaches the equilibrium value k_d. The value of the partition coefficient depends on the radionuclide and the physical and geochemical conditions of the system particulate matter–water.

From Eqs. 19 and 20 we obtain that, in the regions of the drainage area where k_d, θ, and δ are constant, the contaminant concentration is governed by the following equation:

$$\frac{\partial C}{\partial t} = \mathbf{div}\ \mathbf{K}^* \times \mathbf{grad}C - \mathbf{div}\ \mathbf{V}^*\ C - \lambda C \tag{21}$$

where

$$\mathbf{K}^* = \frac{\mathbf{K}}{R} \tag{22}$$

and

$$\mathbf{V}^* = \frac{\mathbf{V}}{R} \tag{23}$$

R is the so-called retardation factor:

$$R = 1 + k_d\frac{\delta}{\theta} \tag{24}$$

The migration of radionuclides is “retarded” by the processes of interaction with particulate matter. Indeed, the components of the effective velocity and of the dispersion tensor $\mathbf{V}^*$ and $\mathbf{K}^*$ in the two phases system water–particulate matter are lower than the corresponding values of $\mathbf{V}$ and $\mathbf{K}$ in water (R>1). This effect is more marked for contaminants characterized by higher values of k_d.

Equation 19 can be also used to predict the radionuclide migration in the over land waters flowing through the catchments provided that θ is assumed approximately equal to 1 m^3 m^{-3} and the dissolved concentration C of radionuclide in Eq. 18 to calculate flux $\mathbf{F}$ is replaced by the total radionuclide concentration (dissolved+particulate form).

As previously noted, the above described approach is based on the essential hypothesis of a quick and complete equilibrium between the dissolved and the attached phases. Furthermore, it is assumed that, in spite of the heterogeneous composition of the substrate, the use of an average value of the partition coefficient is sufficiently appropriate for assessing the ratio between the radionuclide phases. Unfortunately, such hypotheses underlying Eq. 20 are not generally held. The equilibrium is not instantaneously achieved and, moreover, the adsorption–desorption processes are not ever reversible and may depend on the characteristics of the different components of the substrate.

Many authors have discussed and analyzed the dynamic behavior of sorption–desorption processes of dissolved pollutants with sediments and solid matter [35]. Several multiphase dynamics models accounting for complex sorption–desorption kinetics and nonreversible interaction processes have been developed [36].

As formulae (22), (23), and (24) show, when the interaction of a radionuclide with particulate matter is negligible (very low values of k_d as in the case of ^{3}H) the retardation factor R is approximately equal to 1 and the migration velocity and the dispersion coefficient of the contaminant are equal to the corresponding values in Eqs. 16 and 17.

To model the transport of a contaminant by the water flowing through a catchment is complicated by the necessity of accounting for the complex geometry of channels and fractures in the rocks. Moreover, at least in principle, the sub-catchment should be subdivided in "cells" each characterized by specific geological, geochemical, and environmental properties that influence the contaminant migration.

Frequently, discretization of the model equations is performed along the vertical axis. Many models account for several compartments corresponding to different soil layers to assess the vertical migration of radionuclides. A typical multi-compartment model was proposed by Korhonen [37], in which soil in the drainage area was subdivided into layers of various thicknesses. The migration of radionuclides was evaluated accounting for the fluxes of infiltrating water.

However, as experienced for many other kinds of models, increasing the level of detail in the description of the processes does not guarantee greater accuracy of model performance [29]. Indeed, the overall uncertainty of a model is significantly influenced by the contribution of nonnegligible uncertainties from a large number of parameters whose values cannot be known with a sufficient accuracy at site-specific level [38].

It is commonly assumed that two main factors can affect the reliability of a model for predicting the behavior of complex systems:

- The lack of detailed knowledge at the level of the actual processes
- The lack of parameter values necessary for the quantitative assessment of the processes

It has been frequently argued that reductionistic models require a vast amount of data and information and, consequently, may be difficult to apply in practical circumstances. Nevertheless, the reductionistic approach provides a theoretical basis for understanding the role of the fundamental processes involved in the migration of radionuclides through the environment.

Reductionistic models have the undoubted merit of framing the available knowledge concerning the behavior of radionuclides in catchments in a coherent system of logical assumptions, helping the understanding of the functioning of the complex process of contaminant migration from drainage areas.

Principles Underlying Aggregated Models

The alternative strategy for developing models to predict the behavior of contaminants in complex environmental systems is based on the aggregation of processes and parameters. Aggregated models assume that the many different processes of migration of radioactive substances from catchments can be grouped in terms of radionuclide storage, retention, and release by spatially distributed components that, in spite of the heterogeneity of their constituent parts, can be assumed as functional units playing specific roles in the overall migration process.

The radionuclide flux to the water body is a function of time and depends on the rate of deposition of contaminant onto the catchment surface. One of the most simple (or, probably, the most simple) aggregated model was described by Carlsson [39]. Such a model was developed by accounting for the empirical information obtained before the Chernobyl accident. The approach assumes that an initial fraction $k_1 D(t)$ of radionuclide deposited $(D(t))$ per unit time and per unit surface at instant t is instantaneously transferred by runoff waters to the water body (k_1 is dimensionless). Thereafter, the remaining part of deposited radionuclide accumulates in some catchment compartments, such as soils, snowpack, glaciers, etc., that act as "storage units." The amount $S(t)$ of radionuclide per square meter in the storage compartment is available for delayed release and is washed off at a rate $k_2S(t)$ where k_2 (s^{-1}) is a multiplicative factor. Therefore, the flux of radionuclide from the catchment per square meter, $F_R(t)$ ($Bq\,m^{-2}s^{-1}$), can be calculated as follows:

$$F_R(t) = k_1\,D(t) + k_2\,S(t) \tag{25}$$

The model (Fig. 5) corresponds to the conceptual scheme adopted by Yamagata and coworkers [5] that accounts for the fraction of the deposition that is instantaneously transferred to the runoff water and for the removal of radionuclide from the ground deposit.

The coefficient k_2 corresponds to λ_w in Eq. 3. Estimates of k_1 and k_2 were obtained from experimental data available for chronic releases (the fallout from the nuclear weapon tests in the atmosphere). The results of a review by Helton and coworkers [40] suggested that the values of k_1 varied from 0.5×10^{-2} to 12.2×10^{-2} and from 0.1×10^{-2} to 1.9×10^{-2} for ^{90}Sr and ^{137}Cs, respectively. Estimates of k_2 ranged from 2.2×10^{-11} s^{-1} to 1.0×10^{-9} s^{-1} for ^{90}Sr and from 2.1×10^{-12} s^{-1} to 1.8×10^{-10} s^{-1} for ^{137}Cs. For plutonium isotopes, the order of magnitude of k_2 was 10^{-11} s^{-1}.

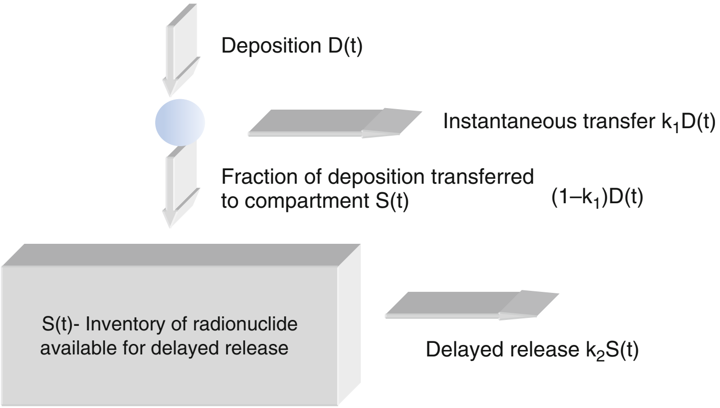
It is possible to derive the following differential equation controlling the time behavior of $S(t)$ from the radionuclide balance in the catchment:

$$\frac{dS(t)}{dt} = -(\lambda + k_2)S(t) + (1 - k_1)D(t) \tag{26}$$

where λ is the radioactive decay constant (s^{-1}).

It is worthwhile to note that k_2 in Eq. 26 corresponds to λ_2 in Eq. 10 as can be easily realized by solving Eq. 26 for a pulse deposition and accounting for Eq. 25.

Unfortunately, as previously noted, before the Chernobyl accident, the contamination of large catchments was essentially caused by chronic releases of radionuclides from the weapon nuclear tests in the atmosphere. This made particularly difficult the evaluation of the parameters in the proposed models. For instance, the estimated values of removal rates (k_2) for strontium are higher that the corresponding values for cesium. Therefore, according to these estimates, the concentration of radiostrontium in runoff water should decline faster than the concentration of radiocaesium. This is in disagreement with the experimental evidence as the empirical evaluations of the TF parameters reported in Table 2 show (note that the values of λ_2 for ^{137}Cs are significantly higher than the corresponding values for ^{90}Sr). Moreover, the effective decay rates of ^{137}Cs concentration in water determined following the Chernobyl accident range from


Deposition D(t)
Instantaneous transfer k_1D(t)
Fraction of deposition transferred to compartment S(t)
(1–k_1)D(t)
S(t)- Inventory of radionuclide available for delayed release
Delayed release k_2S(t)

Radionuclide Migration from Catchments, Modeling. Figure 5
Structure of a simple model for predicting the migration of radionuclides from catchments

3.6×10^{-9} to 2.7×10^{-8} s^{-1}. These values are significantly higher than those reported in the studies performed before the accident [5, 40].

The chronic contamination caused by radionuclide fallout from nuclear weapon tests in atmosphere prevented an accurate evaluation of the time behavior (and consequently of the effective decay parameters) of radionuclides in water. Moreover, these assessments were based on the assumption that radionuclides accumulated in the catchment storage compartments $S(t)$ were fully available for migration. This hypothesis implies a significant underestimate of k_2 when such a parameter is evaluated from the radionuclide balance in the catchment by accounting for the total deposit and for the radionuclide washout as it was performed by pre-Chernobyl assessments.

A Technique to Derive the Radionuclide Migration Model from the Transfer Function The mathematical characteristics of the radionuclide transfer function (10) derived from the analysis of the available empirical data suggest a technique to develop an empirically based aggregated model for assessing the migration of radionuclides from catchments.

Before embarking upon complicated technical matters, it is useful to remind some mathematical notions that are of importance for the following discussion.

Let $GF(t, \tau)$ be the flux (Bq s^{-1}), at time t, of contaminant from the catchment to a water body following a single unit pulse (1 Bq m^{-2}) of radionuclide deposition at instant τ. Let $\delta(t - \tau)$ denote the so-called Dirac's delta function, representing a unit pulse function at time τ. Following a deposition pulse $D \cdot \delta(t - \tau)$, the amount of contaminant $G(t, \tau)$ (Bq s^{-1}) flowing from the catchment is:

$$G(t, \tau) = D \cdot GF(t, \tau) \tag{27}$$

Equation 27 simply states that the radionuclide flux is proportional to the deposition (it is assumed that the process is linear). A chronic deposition $D(t)$ (Bq m^{-2}s^{-1}) can be interpreted as a continuous series of pulses $D(\tau) \cdot \delta(t - \tau)$ occurring at infinitesimal intervals of time dτ. Consequently, $G(t)$, the radionuclide flux from the catchment due to the deposition rate $D(t)$, can be obtained by integrating (27) over the domain of τ (we assume that $D(t)$ is different from 0 when $t > 0$):

$$G(t) = \int_0^t GF(t, \tau) \cdot D(\tau) \mathrm{d}\tau \tag{28}$$

Equation 28 shows that it is possible to determine the radionuclide flux from a catchment for any chronic deposition rate $D(t)$ if function $GF(t, \tau)$ is known. In principle, $GF(t, \tau)$ definitely characterizes all the features of the dynamics of the radionuclide migration process that are necessary for assessing the pollutant flux from a contaminated catchment. It should be noted that $GF(t, \tau)$ corresponds to the so-called transfer function $F(t)$ that was defined in section "Experimental Results Following the Chernobyl Accident." In other words, $F(t)$ should be interpreted as an "empirical estimate" of $G(t, \tau)$ for $\tau = 0$ (the instant of radionuclide pulse deposition).

As previously described, the analysis of contamination data collected following the Chernobyl accident showed that the measured radionuclide flux from a catchment can be fitted to Eq. 10. Consequently, accounting for Eq. 28 and for the equivalence between $GF(t,\tau)$ and $F(t)$ it is possible to write:

$$G(t) = \varepsilon \int_0^t D(\tau) \sum_i \Phi(t)^{\alpha_i} \beta_i A_i \mathrm{e}^{-(\lambda_i + \lambda)(t - \tau)} \mathrm{d}\tau \tag{29}$$

(as usual, we have assumed that the deposition rate $D(t)$ is different from 0 for $t > 0$). In principle, $GF(t, \tau)$ depends on the specific characteristic of the catchment. However, the empirical results previously presented suggest that this function shows a general mathematical form though the values of the parameters in Eq. 29 can be site specific. Putting:

$$S_i(t) = \int_0^t A_i e^{-(\lambda_i + \lambda)(t - \tau)} D(\tau) dt \tag{30}$$

it is possible to write:

$$G(t) = \varepsilon \sum_i \Phi(t)^{\alpha_i} \beta_i S_i(t) \tag{31}$$

It can easily be shown that S_i are solutions of the following system of differential equations:

$$\frac{\mathrm{d}}{\mathrm{d}t} S_i(t) = -(\lambda_i + \lambda) S_i(t) + A_i D(t) \tag{32}$$

with the initial conditions $S_i(0)=0$. Although Eqs. 31 and 32 were obtained from Eq. 29 by a mere mathematical artifice, S_i can be interpreted as the amount of radioactive substance ($\mathrm{Bq\,m^{-2}}$) accumulated in the ith radionuclide storage compartment of the catchment and available for migration to the water body. λ_i account for all those processes, such as the radionuclide fixation to soil, that makes the removal of the radioactive substances from the catchment by the runoff waters less effective. Parameters A_i are the fractions of deposited radionuclide accumulated in the ith compartment and the products $\varepsilon\Phi(t)^{\alpha_i}\beta_i(\mathrm{m^2\,s^{-1}})$ are the transfer factors of radionuclide from the ith compartment to the runoff water at instant t. It is worthwhile to note that the instantaneous transfer of deposited radionuclide in the Carlsson–Yamagata model corresponds to a storage compartment characterized by a value of λ_i that tends to infinity.

According to Håkanson [41], two main storage compartments correspond to the fast and slow components in the transfer function (10): (a) a fast component – wetlands dominated by a fast turnover of substances and horizontal migration processes; (b) a slow component – dryland dominated by vertical migration processes, through the soil horizons, and by the ground water dispersion.

The described aggregated model is based on the assumption that ε, A_i, and β_i are independent of time. In principle, alternative formulations are possible in view of the seasonal effects that can influence the migration of radionuclide from catchments and, therefore, the values of the model parameters. It is important to note that the available empirical information refers, almost exclusively, to the environmental contamination caused by the Chernobyl accident that occurred during spring when particular and typical seasonal processes like snow and ice melting occurred.

Theoretical Basis of "Aggregated" Models The amount F_R ($\mathrm{Bq\,s^{-1}}$) of radionuclide removed, per unit time, by the runoff waters from a catchment is:

$$F_R = A\phi C \tag{33}$$

where A is the surface of the catchment ($\mathrm{m^2}$), ϕ is the yearly average flux of water per square meter flowing through the catchment ($\mathrm{m^3\,s^{-1}\,m^{-2}}$) and C, as usual, is the concentration of contaminant in water. ϕ is the sum of two components: the infiltrating water (ϕ_2) and the surface runoff (ϕ_1). Due to the relatively high values of k_d and, consequently, of the retardation factors of radionuclides such as ^{137}Cs and ^{90}Sr, it is reasonable to assume that ϕ_2 does not contribute significantly to the migration of the contaminant to the water body.

The time behavior of the radionuclide inventory per square meter, I ($\mathrm{Bq\,m^{-2}}$), in the catchment is controlled by the following first-order differential equation:

$$A\frac{\mathrm{d}I}{\mathrm{d}t} = -F_R - \lambda \cdot A \cdot I \tag{34}$$

Let K ($\mathrm{m^{-1}}$) be the ratio of the radionuclide concentration in water divided by the radionuclide inventory:

$$K = \frac{C}{I} \tag{35}$$

K is the so-called entrainment coefficient used by several authors [27, 42, 43] to investigate, by quantitative approaches, the migration of radionuclides from catchments.

The radionuclide activity balance requires that:

$$\frac{\mathrm{d}I}{\mathrm{d}t} = -\phi \cdot K \cdot I - \lambda \cdot I \tag{36}$$

thus

$$I = I_0\,\mathrm{e}^{(-\phi K t - \lambda t)} \tag{37}$$

and, from Eqs. 33 and 35,

$$F_R = I_0\,A\phi K\mathrm{e}^{(-\phi K t - \lambda t)} \tag{38}$$

where I_0 is the value of I at instant $t=0$.

The radionuclide flux, F_1, due to the surface runoff is:

$$F_1 = I_0\rho A\phi K\mathrm{e}^{(-\phi K t - \lambda t)} \tag{39}$$

where ρ is ϕ_1/ϕ.

In particular, the concentration in the surface runoff water is:

$$C = I_0\,K\mathrm{e}^{(-\phi K t - \lambda t)} \tag{40}$$

The above results were obtained from a simple application of a general notion of entrainment coefficient K

without any particular assumption concerning the mechanisms of interaction of the radionuclide with the different kinds of rock and soil substrata in the catchment (however, K should be assumed independent of time).

To fix the ideas it can be useful to evaluate K under the hypothesis that the partition coefficient approach can be applied to determine the radionuclide concentration in water [44]. The contaminant transported through the catchment by the surface runoff is, prevailingly, confined in an upper layer of soil of thickness ζ (m). Such a hypothesis is valid when the contaminant strongly interacts with the soils and the rocks in the catchment as it occurs for radionuclides characterized by high values of the partition coefficient k_{d}.

The inventory I may be calculated as follows:

$$I = C\zeta(\theta + \delta k_d) \tag{41}$$

where θ ($\mathrm{m}^3\mathrm{m}^{-3}$) is the water content of soil and δ ($\mathrm{kg\,m}^{-3}$) is the soil density. As for the mentioned radionuclides:

$$\delta k_d >> \theta \tag{42}$$

from formula (35), it follows:

$$K = \frac{1}{\zeta\delta k_{\mathrm{d}}} \tag{43}$$

It is worthwhile to note that the flux of radionuclide F_1 is expressed by a single exponential function when values of the parameters ρ, ϕ, and K averaged over the whole catchment are used in Eq. 39. This result is in disagreement with the empirical data available following the Chernobyl accident. Indeed, Figs. 3 and 4 clearly show that at least two exponential components can be observed over a period of several years. Moreover, since K is proportional to the reciprocal of k_{d}, the calculated values (Eq. 43) of the entrainment coefficient for Cs are significantly lower than the corresponding values for Sr. Consequently, Eq. 40 predicts a faster decline of the concentration in water of the latter radionuclide in disagreement with the empirical evidence as already noted in section "Principles Underlying Aggregated Models."

However, the characteristics that control the migration of a radionuclide through a catchment can be very variable on a spatial scale. For instance, the values of the distribution coefficient depend on the soil characteristics and on the chemical properties of the pollutant. Of course, in a complex catchment, the values of k_{d} for a particular pollutant may vary within a large range in view of the spatial variability of the geological properties of the soils. The assessment of the flux of radionuclide transported by runoff waters requires that the catchment is subdivided in subregions, each of which is characterized by particular values of ρ, ϕ, and K. The overall migration process can be modeled by integrating the fluxes of radionuclide from each subcatchment over the whole drainage area.

Putting $\phi K = \omega$, the pollutant flux can be evaluated as follows:

$$F_1 = I_0 A\bar{\rho}\mathrm{e}^{(-\lambda t)}\int_0^\infty \omega f(\omega)\mathrm{e}^{(-t\omega)}\mathrm{d}\omega \tag{44}$$

where $f(\omega)$ is the distribution function of ω ($\bar{\rho}$ is the average of ρ over the catchment). Equation 44 is valid if ρ and ω are statistically independent.

The effective decay rate, $\lambda^*(t)+\lambda$, of radionuclide concentration in water at time t is:

$$\lambda^*(t) + \lambda = -\frac{1}{C}\frac{\mathrm{d}C}{\mathrm{d}t} = -\frac{1}{F_1(t)}\frac{\mathrm{d}F_1(t)}{\mathrm{d}t} \tag{45}$$

$\lambda^*(t)$ controls the decay of radionuclide concentration in runoff water due to environmental effects.

After boring but easy calculations and an integration by parts, we get:

$$\lambda^*(t) = \frac{1}{t} + \frac{1}{t}\frac{\int_0^\infty \omega g'(\omega)\mathrm{e}^{(-t\omega)}\mathrm{d}\omega}{\int_0^\infty g(\omega)\mathrm{e}^{(-t\omega)}\mathrm{d}\omega} \tag{46}$$

where $g(\omega) = \omega \cdot f(\omega)$. The second term in the right hand side of the previous equation may scarcely affect the order of magnitude of $\lambda^*(t)$ due to the presence of $g'(\omega)$ in the integral at numerator. Such a derivative reaches, indeed, both positive and negative values that may mutually compensate when the function is integrated over $(0, \infty)$.

$\lambda^*(t)$ can be evaluated for some simple distribution functions. Several examples show that such a function depends prevailingly on the observation time t but may be less sensitive to the environmental conditions and to the pollutant characteristics [17, 45].

If we hypothesize that ω is log-normally distributed:

$$f(\omega) = \frac{1}{\sigma\sqrt{2\pi}}\frac{1}{\omega}\exp\left[-\frac{1}{2\sigma^2}(\ln\omega-\mu)^2\right] \tag{47}$$

we obtain:

$$\lambda^*(t) = \frac{\phi(t)}{t} \tag{48}$$

where

$$\phi(t) = 1 + \frac{1}{\sigma^2}\left(\mu - \frac{\int_0^\infty \exp\left(-\frac{(\ln\omega-\mu)^2}{2\sigma^2}\right)\ln\omega\exp(-\omega t)\mathrm{d}\omega}{\int_0^\infty \exp\left(-\frac{(\ln\omega-\mu)^2}{2\sigma^2}\right)\exp(-\omega t)\mathrm{d}\omega}\right) \tag{49}$$

μ and σ are, respectively, the mean value and the standard deviation of the distribution of the natural logarithm of ω. $\phi(t)$ is close to 1 when σ is very large ($\lim_{\sigma\to\infty}\phi(t)=1$).

The statistical aggregation of the contaminant migration fluxes from different points of a catchment can lead to slight variability in the dynamics of the overall migration process. Equation 48 shows that the value of $\lambda^*(t)$ depends on the observation time. In particular, the average value of $\lambda^*(t)$ over an interval of time centered around t should be of the order of $\langle\phi(t)/t\rangle$ (the angular brackets denote the time average).

Function (49) can be evaluated for ^{137}Cs and ^{90}Sr using the experimental log-normal distribution of k_d. For values of the parameters in Eq. 43 and in the distribution function (47) commonly measured for Cs and Sr in soils [46], $\phi(t)$ is almost independent of time (at least when t is of the order of months or years). Therefore, it is possible to write:

$$\lambda^*(t) \approx \frac{k}{t} \tag{50}$$

where k is a dimensionless constant of the order of 1. The previous results explain several characteristics of the time behavior of radionuclide concentration in runoff waters. Indeed, the experimental evaluations of λ_1 and λ_2 in Table 2 show that, following a single pulse of radionuclide deposition, the values of such parameters are approximately of the order of the reciprocal of t (the time since the accidental release of radionuclide occurred) averaged over the period of observation (months for the first exponential component and few years for the second) in agreement with Eq. 50. The empirical values of the effective environmental decay rates λ_1 and λ_2 (Table 2) are slightly variable despite the large range of characteristics of the catchments and the significantly different chemical properties of radionuclides like ^{137}Cs and ^{90}Sr. For instance, the values of λ_2 for ^{90}Sr are less than one order of magnitude lower than the corresponding values for ^{137}Cs, whereas the values of k_d for these radionuclides differ up to several orders of magnitude. Note that $\langle\lambda^*(t)\rangle$ is an estimate of λ_1 and λ_2 when the average is calculated over the above mentioned periods of time.

All the previous calculations are based on several assumptions and simplifications (for instance, ϕ_1, the yearly average surface runoff, is independent of time). However, the conceptual approach of the model sheds light on the role played by sub-catchments characterized by different radionuclide entrainment properties. It seems reasonable that, around a given instant t, the catchment areas characterized by values of the product ϕK of the order of $1/t$ contribute more effectively to the radionuclide migration. Indeed, the contributions from areas with $\phi K >> 1/t$ are negligible in view of the high value of the exponent in Eq. 38, whereas, if $\phi K << 1/t$, the contribution is negligible in view of the low value of the multiplicative factor ϕK in the equation.

Modeling Particular Processes of Radionuclide Migration from Catchments

Analyses based on process and system aggregations at level of the whole catchment can be useful as it has been shown in the previous sections. However, it is essential to define the most appropriate spatial scale and resolution time in view of the applications of this kind of models.

It is important to note that the described aggregated models are essentially aimed at assessing the radionuclide fluxes in hydrological conditions representing the average situation throughout the whole catchment. For a reliable and proper application of the models, the spatial resolution should be based on the identification of the areas in the catchment where particular hydrologic processes may significantly influence the radionuclide migration. For instance, the role of river floodplains should be, in some circumstances, carefully accounted for in view of radionuclide removal and remobilization when such areas are significantly contaminated. For example, the peak concentration of ^{90}Sr in water of Prypiat River showed in Fig. 4 at instant "t = 62nd month" corresponds to the inundation, in January 1991, of the heavily polluted Prypiat floodplain in the vicinity of the Chernobyl Nuclear Power Plant. The flooding was caused by an "ice jam" formed in the Prypiat River channel between the Yanov bridge and the town of Chernobyl. The consequent washout of the radionuclide from the flooded area into the river water resulted in a significant increase of ^{90}Sr concentration [47] that cannot be properly predicted by Eq. 11 and by the corresponding model (section "A Technique to Derive the Radionuclide Migration Model from the Transfer Function") when the aggregation is performed at level of the whole catchment of the river.

Further, sub-catchment areas covered with snow and glaciers play a particular role in the radionuclide transfer process. For instance, it was found [48] that, in connection with the high water fluxes caused by ice and snow melting, the concentrations of dissolved cesium in water can significantly increase. For example, in spring, the concentrations of ^{137}Cs of Chernobyl origin in the water of the Dora Baltea River, whose catchment in Northwest Italy comprises high mountain areas in the Alps, were at least one order of magnitude higher than the concentrations measured in the other seasons. On the contrary, the high water fluxes in the river following the heavy precipitation in autumn were not associated with a similar corresponding increase of dissolved radionuclide concentration in water. A direct application of Eq. 11 is not appropriate to predict such a behavior as the modeled levels of radionuclide concentration in water are not correlated to the water flux in the river due to the low values of α_2. However, in order to properly predict the different radionuclide contributions associated with the sub-catchments covered by ice and snow, the catchment of the river can be subdivided in subregions characterized by different values of the transfer coefficient ε. The values of ε in the sub-catchments covered by ice and snow should be assumed significantly higher than the corresponding values in the other areas of the drainage basin. With these assumptions, the different levels of dissolved radionuclide concentration in river water during the flow peaks in spring and autumn can be easily explained by the different proportions of the rain runoff and of the melting waters flowing from areas covered by ice and snow.

Conclusions

The migration of radionuclides from the terrestrial to the aquatic environment involves complex processes of hydrological, physical, and geochemical nature. In principle, the models aimed at predicting the transfer of radioactive substances from catchments to water bodies can be classified as reductionistic and aggregated. However, most of the models developed for practical applications are hybrids that profit from both techniques to assess the complex behavior of radionuclide migrating from drainage areas.

Reductionistic models are aimed at predicting the migration of radionuclides according to primary laws from fundamental disciplines such as physics and chemistry and by accounting for as many detailed processes as reasonably possible. On the contrary, aggregated models are based on the identification of environmental units and processes whose behaviors can be explained by laws that are derived from empirical observations and that concern each unit as a whole (this could justify the term "holistic" occasionally used to define aggregated models). It is important to emphasize that, in this entry, the word "aggregation" refers to a conceptual empirical approach rather than to the mathematical techniques commonly used to simplify the structure of a model.

The advantages and the disadvantages of reductionistic and aggregated models for predicting the migration of radionuclides through the aquatic environment have been extensively analyzed and discussed in the scientific literature chiefly in view of applications for the management of environmental emergencies [26–28, 49–51]. It is generally recognized that reductionistic models have the merit of framing in a rational scheme the many processes controlling the transfer of radionuclides from catchments to water bodies and, therefore, are helpful for understanding the different roles of the migration mechanisms and their reciprocal interactions. Aggregation is a useful strategy for developing models that can be easily applied for practical purposes and especially in emergence situations where model input data is limited without extensive collection of empirical data in the field.

Future Directions

A Discussion Including Potential Impacts on the Development of Certain Areas of Science

Radioecology and radiological modeling are traditional scientific disciplines that have been the subject of a great many studies carried out during the past decades. Pulse-type accidents of environmental contamination at continental scale are infrequent and the experiences gained following the Chernobyl accident are somewhat unique in environmental sciences. It is reasonable to think that the results from these experiences can be useful for addressing similar problems concerning the migration from catchments of other kinds of pollutants such as heavy metals.

As for many other radioecological studies, the investigations of the migration of radionuclide from catchments have produced several consolidated results that are widely accepted by most experts. However, some further model improvements are still necessary. The available aggregated models have been applied and validated chiefly for two radionuclides, ^{137}Cs and ^{90}Sr, that can significantly contribute to radiological doses to man from the aquatic pathway as demonstrated by the environmental consequences of the Kysthym and the Chernobyl accidents. It can be of interest, however, to extend the investigations to other radionuclides in view of other kinds of accidents causing releases of radioactive substances into the environment.

The analysis of the mathematical structure of the transfer function raises a variety of questions relevant to the quantitative assessment of radionuclide migration from catchments, such as the dependence of the function coefficients from the seasonal conditions during accidental releases. Further, the possible occurrence of nonlinearity effects caused by the water flux deserves to be better investigated, chiefly in relation to the early period following an accident.

Both reductionistic and aggregated models require calibration and customization for the applications to the specific catchments of interest. This can entail a significant effort for complex reductionistic models. It is desirable that appropriate strategies are developed in the future to solve such complex problems.

In general, models to simulate radionuclide migration through and from catchments are available. However, a major difficulty is the lack of suitable model input data at the appropriate scale. A profitable strategy for developing models aimed at assessing the radionuclide behavior in catchments is based on the identification of emerging processes and on their parameterization accounting for the specific environmental conditions and situations that influence migration processes.

Bibliography

Primary Literature

1. Horowitz AJ, Lum KR, Garbarino JR, Hall GEM, Lemieux C, Demas CR (1996) Problems associated with using filtration to define dissolved trace element concentrations in natural water samples. Environ Sci Technol 30:954–963
2. Eisenbud M (1963) Environmental radioactivity. McGraw-Hill, New York/San Francisco/Toronto/London
3. Fowler EB (1965) Radioactive fallout, soils, plants, foods, man. Elsevier, Amsterdam/London/New York
4. Menzel RG (1960) Transport of strontium-90 in runoff. Science 13:499–500
5. Yamagata N, Matsuda S, Kodaira K (1963) Run-off of caesium-137 and strontium-90 from rivers. Nature 200:668–669
6. De Cort M, Dubois G, Fridman D, Germenchuk M, Izrael YA, Janssen A, Jones AR, Kelly GN, Kvasnikova EV, Matveenko II, Nazarov IM, Pokumeiko YM, Sitak VA, Stukin ED, Tabachny LY, Tsaturov YS, Avdyushin SI (1998) Atlas of caesium deposition on Europe after the Chernobyl accident. EUR Report 16733, Luxemburg
7. Hilton J, Livens FR, Spezzano P, Leonard DRP (1993) Retention of radioactive caesium by different soils in the catchment of a small lake. Sci Total Environ 129:253–266
8. Santschi PH, Bollhander S, Zingg S, Luck A, Farrenkothen K (1990) The self-cleaning capacity of surface waters after radioactive fallout. Evidence from European Waters after Chernobyl, 1986–1988. Environ Sci Technol 24:519–527
9. Monte L (1995) Evaluation of radionuclide transfer functions from drainage basins of fresh water systems. J Environ Radioactiv 26:71–82
10. Kaniviets VV, Voitcekhovich OV (1992) Scientific report: Radioecology of water systems in zone of consequences of Chernobyl accident. Report of Ministry of Chernobyl Affairs of Ukraine. Contract Number 1/92 (in Russian)
11. Mundschenk H (1992) Ueber Nachwirkungen des Reaktorunfalls in Tschernobyl im Bereich der "alten" Bundeswasserstrassen. (On long term effects of the Accident of the Nuclear Power Plant in Chernobyl in the German

Federal Waterways) Sonderdruck aus: Deutsche Gewaesserkundliche Mitteilungen 36:7–19 (in German)

12. Maringer FJ (1994) Das Verhalten von Radionukliden im Wasser, Schwebstoff und Sediment der Donau. Dissertation, Technischen Universität Wien (in German)
13. Sunblad B, Bergström U, Evans S (1991) Long term transfer of fallout from the terrestrial to the aquatic environment. In: Moberg L (ed) The Chernobyl fallout in Sweden. Results from a research program on environmental radiology. The Swedish Radiation Protection Institute, Stockholm
14. Vray F, Debayle C, Louvat D (2003) Long-term flux of Chernobyl-derived ^{137}Cs from soil to French rivers: A study on sediment and biological indicators. J Environ Radioactiv 68:93–114
15. Smith JT, Clarke RT, Saxén R (2000) Time-dependent behaviour of radiocaesium: a new method to compare the mobility of weapons test and Chernobyl derived fallout. J Environ Radioactiv 49:65–83
16. Smith JT, Konoplev AV, Bulgakov AA, Comans RNJ, Cross MA, Kaminski S, Khristuk B, Klemt E, de Koning A, Kudelsky AV, Laptev G, Madruga M-J, Voitsekhovitch OV, Zibold G (2002) Simplified models for predicting ^{89}Sr, ^{90}Sr, ^{134}Cs, ^{137}Cs, ^{131}I in water and fish of rivers, lakes and reservoirs, AQUASCOPE Technical Deliverable. CEH Centre for Ecology and Hydrology, Natural Environment Research Council, United Kingdom
17. Monte L (1997) A collective model for predicting the long-term behaviour of radionuclides in rivers. Sci Total Environ 2001:17–29
18. Smith JT, Wright SM, Croos MA, Monte L, Kudelsky A, Saxen R, Vakulovsky A, Timms D (2004) Global analysis of the riverine transport of ^{90}Sr and ^{137}Cs. Environ Sci Technol 38:850–857
19. Kivva SL, Zheleznyak MI (2000) Hydrological and physico-chemical processes determining radionuclide redistribution. In: van der Perck M, Svetlitchnyi AA, den Basten J, Wielinga A (eds) SPARTACUS, Spatial redistribution of radionuclides within catchments. Final report EC Contract n° IC15CT980215. Utrecht University, The Netherlands
20. Håkanson L (2006) Suspended particulate matter in lakes, rivers, and marine systems. The Blackburn Press, Caldwell
21. Wanielista PM (1990) Hydrology and water quality control. Wiley, New York
22. Benes P, Picat P, Gernik M, Quinault S (1992) Kinetics of radionuclide interaction with suspended solids in modeling the migration of radionuclides in rivers. (Part I and II). J Radioanal Nucl Chem 159:175–200
23. Cremers A, Henrion PN (1984) Radionuclide partitioning in sediments: theory and practice. In: Seminar on the Behaviour of Radionuclides in Estuaries, Renesse (The Netherlands), 17–21 September 1984. Commission of the European Communities, XII/380/85-EN: 1–25
24. Delle Site A (2000) Factors affecting sorption of organic compound in natural sorbent/water systems and sorption coefficients for selected pollutants. A review. J Phys Chem Ref Data 29(6):1–253
25. Zheleznyak M, Demchenco R, Khursin S, Kuzmenko Yu, Tkalich P, Vitjuk N (1992) Mathematical modelling of radionuclide dispersion in the Prypiat-Dnieper aquatic system after the Chernobyl accident. Sci Total Environ 112:89–114
26. Garcia-Sanchez L (2008) Watershed wash-off of atmospherically deposited radionuclides: review of the fluxes and their evolution with time. J Environ Radioactiv 99:563–573
27. Garcia-Sanchez L, Konoplev AV (2009) Watershed wash-off of atmospherically deposited radionuclides: a review of normalized entrainment coefficients. J Environ Radioactiv 100:774–778
28. Monte L, Brittain EJ, Håkanson L, Smith JT, van der Perk M (2004) Review and assessment of models for predicting the migration of radionuclides from catchments. J Environ Radioactiv 75:83–103
29. Håkanson L, Monte L (2003) Radioactivity in Lakes and Rivers. In: Scott EM (ed) Modelling radioactivity in the environment. Elsevier Science, Oxford
30. Agüero A, García-Olivares (2000) CIEMAT model results for Esthwaite water. In: Modelling of the transfer of radiocaesium from deposition to lake ecosystems. Report of the VAMP Aquatic Working Group. IAEA-TECDOC-1143, International Atomic Energy Agency, Vienna
31. Bäverstam U, Fraser G, Kelly GN (eds) (1997) Decision making support for off-site emergency management. Radiat Prot Dosim 73:1–317
32. BIOMOVS II (1996) Wash-off of Sr-90 and Cs-137 from two experimental plots: model testing using Chernobyl data. Technical report No. 9. Swedish Radiation Protection Institute, Stockholm
33. Heling R, Zheleznyak M, Raskob W, Popov A, Borodin R, Gofman D, Lyashenko G, Marinets A, Pokhil A, Shepeleva T, Tkalich P (1997) Overview of the modelling of hydrological pathways in RODOS. Radiat Prot Dosim 73:67–70
34. International Atomic Energy Agency (IAEA) (2010) Handbook of parameter values for the prediction of radionuclide transfer in terrestrial and freshwater environments. Technical Reports Series no. 472. Vienna
35. Karickoff SW (1986) Pollutant sorption in environmental systems. In: Brock Nelly W, Blau GE (eds) Environmental exposure from chemicals, vol I. CRC Press, Boca Raton
36. Comans RNJ, Hockley DE (1992) Kinetics of cesium sorption on illite. Geochim Cosmochim Acta 56:1157–1164
37. Korhonen R (1990) Modeling transfer of ^{137}Cs in a large Finnish watercourse. Health Phys 59:443–454
38. Håkanson L (1995) Optimal size of predictive models. Ecol Modell 78:195–204
39. Carlsson S (1978) A model for the movement and loss of ^{137}Cs in a small watershed. Health Phys 34:33–73
40. Helton JC, Muller AB, Bayer A (1985) Contamination of surface – water bodies after reactor accidents by the erosion of atmospherically deposited radionuclides. Health Phys 48:757–771
41. Håkanson L (2004) A new generic sub-model for radionuclide fixation in large catchments from continuous and single-pulse fallouts, as used in a river model. J Environ Radioactiv 77:247–273
42. Bulgakov A, Konoplev A, Popov V, Scherbak A (1991) Removal of long-lived radionuclides from the soil by surface runoff near the Chernobyl nuclear power station. Sov Soil Sci 23:124–131

43. Konoplev A, Bulgakov A, Popov V, Bobovnikova TI (1992) Behaviour of long-lived Chernobyl radionuclides in a soil–water system. Analyst 117:1041–1047
44. Joshi SR, Shukla BS (1991) The role of water/soil distribution coefficient in the watershed transport of environmental radionuclides. Earth Planet Sci Lett 105:314–318
45. Monte L (1998) Predicting the migration of dissolved toxic substances from catchments by a collective model. Ecol Modell 110:269–279
46. Sheppard MI, Thibault DH (1990) Default soil solid/liquid partition coefficients, k_ds, for four major soil types: a compendium. Health Phys 59:471–482
47. Monte L, Periañez R, Kivva S, Laptev G, Angeli G, Barros H, Zheleznyak M (2006) Assessment of state-of-the-art models for predicting the remobilisation of radionuclides following the flooding of heavily contaminated areas: the case of Pripyat River floodplain. J Environ Radioactiv 88: 267–288
48. Spezzano P, Bortoluzzi S, Giacomelli R, Massironi L (1994) Seasonal variations of ^{137}Cs activities in the Dora Baltea River (Northwest Italy) after the Chernobyl Accident. J Environ Radioactiv 22:77–88
49. Monte L, Brittain JE, Håkanson L, Heling R, Smith JT, Zheleznyak M (2003) Review and assessment of models used to predict the fate of radionuclides in lakes. J Environ Radioactiv 69:177–205
50. Monte L, Boyer P, Brittain JE, Håkanson L, Lepicard S, Smith JT (2005) Review and assessment of models for predicting the migration of radionuclides through rivers. J Environ Radioactiv 79:273–296
51. Monte L, Håkanson L, Periañez R, Laptev G, Zheleznyak M, Maderich V, Angeli G, Koshebutsky V (2006) Experiences from a case study of multi-model application to assess the behaviour of pollutants in the Dnieper-Bug Estuary. Ecol Modell 195:247–263

Books and Reviews

Moldan B, Černý B (eds) (1994) Biogeochemistry of small catchments: a tool for environmental research. SCOPE 51. Wiley, England, pp 419

Sir Warner F, Harrisson RM (eds) (1993) Radioecology after Chernobyl. Biogeochemical pathways of artificial radionuclides. SCOPE 50. Wiley, England, pp 367

Scott EM (ed) (2003) Modelling radioactivity in the environment. Elsevier Science, Oxford, p 427

Degens ET, Kempe S, Richey JE (eds) (1991) Biogeochemistry of major world rivers. SCOPE 42. Wiley, England, pp 356

Wanielista M (1990) Hydrology and water quality control. Wiley, New York, p 565

Håkanson L (2006) Suspended particulate matter in lakes, rivers, and marine systems. Blackburn Press, Caldwell, p 319

Shukla SB (1993) Watershed, river, and lake modeling through environmental radioactivity. Environmental Research & Publications Inc, Hamilton, p 227

Radionuclides as Tracers of Ocean Currents

Pavel P. Povinec[1], Katsumi Hirose[2]
[1]Faculty of Mathematics, Physics and Informatics, Comenius University, Bratislava, Slovakia
[2]Faculty of Science and Technology, Sophia University, Chiyodaku, Tokyo, Japan

Article Outline

Glossary
Definition of the Subject
Introduction
Hydrography Background
Sources of Radionuclides in the Marine Environment
Sampling and Analytical Techniques
Distribution and Transport of Radionuclides in Seawater – Tracing Ocean Currents
Future Directions
Acknowledgments
Bibliography

Glossary

Anthropogenic radionuclides Radionuclides produced by human activities (e.g., in nuclear weapons tests, in nuclear reactors, in particle accelerators).

Global fallout Radioactive contamination from atmospheric nuclear weapons tests, which was widely dispersed in the atmosphere and then deposited on land and sea.

Ocean current A continuous, directed movement of ocean water generated by wind, Coriolis force, temperature and salinity differences, breaking waves, and tides caused by the gravitational pull of the Moon and the Sun.

Ocean gyre A large system of rotating ocean currents associated with large wind movements.

Thermocline A layer in the water column in which temperature changes more rapidly with depth than it does in the layers above or below, separating the upper mixed layer from the calm deep water below.

Thermohaline circulation A large-scale circulation of water masses in the ocean driven by global density gradients created by surface heat and freshwater fluxes.

The southern ocean The southernmost part of the World Ocean, south of 60°S latitude, encircling Antarctica.

Definition of the Subject

Global fallout radionuclides after their main injection from atmospheric nuclear weapons tests on the ocean surface have been used for about 5 decades for studying processes in the marine environment. Conservative (^{3}H, ^{14}C, ^{90}Sr, ^{129}I, ^{137}Cs, and others) as well as nonconservative (^{238}Pu, ^{239}Pu, ^{240}Pu, ^{241}Am, and others) radionuclides have been available for surface-, medium-, and deepwater studies. Using the recent data obtained in the framework of several international projects, it has been possible to study recent trends in the circulation of ocean water masses in the Atlantic, Pacific, and Indian Oceans, as well as interocean exchange. Subtropical gyres, which accumulate radionuclide tracers and other contaminants on timescales of decades, are of environmental importance for the protection of the marine environment from land-based contamination sources.

Introduction

There are four main sources of environmental radionuclides that can be found in the marine environment [1]:

(i) Natural – cosmogenic radionuclides – results of interactions of cosmic rays with atoms in the atmosphere, and their subsequent deposition on the ocean surface (e.g., ^{3}H, ^{7}Be, ^{10}Be, ^{14}C, ^{26}Al, ^{53}Mn).

(ii) Natural – primordial radionuclides (e.g., ^{40}K, ^{238}U, ^{232}Th) and their decay products (e.g., ^{226}Ra, ^{230}Th) found in the earth crust; due to radon emanation its decay products are also found in the atmosphere, and then after deposition in the terrestrial and marine environments (e.g., ^{210}Po, ^{210}Pb).

(iii) Anthropogenic – global fallout radionuclides – produced during atmospheric tests of nuclear weapons (e.g., ^{3}H, ^{14}C, ^{90}Sr, ^{137}Cs, Pu isotopes, ^{241}Am).

(iv) Anthropogenic – radionuclides released from nuclear installations – mostly from reprocessing nuclear facilities (e.g., ^{3}H, ^{14}C, ^{90}Sr, ^{99}Tc, ^{129}I, ^{137}Cs).

This entry concentrates mostly on global fallout radionuclides as they have been most frequently used in oceanic studies. They were injected onto the ocean surface from the atmosphere after large-scale U.S. and former USSR atmospheric nuclear weapons tests carried out mainly during the 1950s and early 1960s [2, 3]. The peak concentrations of the anthropogenic radionuclides in the atmosphere of the Northern Hemisphere were observed in 1963, and with 1 year delay in surface waters of northern oceans and seas [4]. Mapping of this deposition revealed that the major injection on the ocean surface occurred at midlatitudes of the western North Pacific [5]. Global fallout radionuclides in seawater have been measured since the 1960s with the aim to assess radiological impact from nuclear weapons tests [6], and later they were used as transient tracers to investigate oceanographic processes [3, 7–10]. Radionuclides released from nuclear installations, especially from European nuclear reprocessing facilities in Sellafield (UK) and La Hague (France) have also been extensively used for tracing water masses in the eastern North Atlantic ocean and adjacent seas [11–14]. Radioecological studies were carried out to investigate the possible impact of dumped radioactive wastes in the Barents and Kara Sea [15–18], as well as in the Japan Sea and the Northwest Pacific Ocean [19–21]. The collected radionuclide data have been recently stored in marine radionuclide databases (GLOMARD/MARIS [22, 23], www.iaea.org/maris), and HAM [24] radionuclide databases).

There have been two important breakthroughs in the oceanographic research. Firstly, national and/or regional investigations have been shifted to global projects, such as GEOSECS (Geochemical Ocean Sections Study) program, WOCE (World Ocean Circulation Experiment) (www.WOCE.org), or more recently CLIVAR (www.CLIVAR.org) and GEOTRACES (www.GEOTRACES.org). The WOCE program, which was carried out during the 1990s represents the most extensive ^{3}H and ^{14}C sampling carried out in the World Ocean till now. Secondly, new highly sensitive analytical techniques have been implemented. In the case of shorter-lived ^{137}Cs, it was the development of underground facilities for ultra-low-level gamma-ray spectrometry [25–29]. For long-lived radionuclides (e.g., ^{239}Pu, ^{240}Pu), it was a total change in the philosophy of analysis as a shift from counting of radioactive decays (and thus waiting for them) to the direct counting of atoms using mass spectrometry techniques such as Accelerator Mass Spectrometry (AMS) [30–33], Inductively Coupled Plasma Mass Spectrometry (ICPMS) [34–37], Thermal Ionization Mass Spectrometry (TIMS) [38], and Resonance Ionization Mass Spectrometry (RIMS) [39] has been made.

Several radionuclides have extensively been used as tracers in oceanographic studies. Tritium is an ideal tracer because it is directly incorporated into the water molecule. However, its short half-life (12.32 year) restricts its use only for studies of relatively short-term transport processes. ^{3}H is produced either naturally by interactions of cosmic rays with nitrogen and oxygen in the upper atmosphere, but it has also been produced in large amounts in atmospheric nuclear weapons tests. The ^{3}H concentration peaked in the atmospheric moisture in 1963, when it was almost 1,000 times higher than its natural cosmogenic concentration. ^{3}H has also been released in large quantities from nuclear reprocessing facilities, mainly in Sellafield and in La Hague. The penetration of bomb tritium from surface waters into deeper layers of the ocean has been used to study pathways and timescales of deep and bottom water formation [40, 41]. Cosmogenic ^{14}C is produced by cosmic rays in the upper atmosphere, mainly in the reaction ^{14}N(n,p)^{14}C. Its levels in the environment have also been disturbed by bomb-produced ^{14}C, when in 1963 they were by a factor of 2 higher than natural (cosmogenic) levels observed in atmospheric carbon dioxide. Large quantities of ^{14}C have also been produced in nuclear reprocessing facilities. Because of its long half-life (5,730 year), specific stratosphere–troposphere–biosphere interactions, exchange of carbon dioxide between air and surface ocean, and sequestration of carbon dioxide into the ocean interior, ^{14}C became the most frequently studied environmental radionuclide, important for a better understanding of climate change [42, 43]. Oceanic radiocarbon data contributed to a better understanding of thermohaline circulation in the World Ocean known as a Great Ocean Conveyor Belt [44]. Other global fallout radionuclide tracers extensively used in water circulation studies are ^{90}Sr (half-life 28.78 year) and ^{137}Cs (half-life 30.17 year). Both were released in large quantities during atmospheric nuclear weapons tests, as well as from nuclear reprocessing facilities. They are found in seawater mostly in a dissolved phase, following well the movement of water masses. Their removal from the water column is mainly due to their radioactive decay and diffusion [45, 46]. ^{129}I has been introduced to the World Ocean mainly from nuclear reprocessing facilities, and because of its long half-life (15.7×10^{6} year) it represents an alternative tracer to global fallout radionuclides for studies of long-term transport and circulation processes [47]. On the other hand, anthropogenic tracers such as ^{238}Pu (half-life 87.74 year), ^{239}Pu (half-life 2.44×10^{4} year), ^{240}Pu (half-life 6.58×10^{3} year), and ^{241}Am (half-life 458 year) are particle reactive, and due to their chemically reactive nature, significant portions of these radionuclides in the World Ocean have reached marine sediments [8, 48–50].

GEOSECS program [7–9], carried out during the 1970s, was the first large-scale project devoted to investigation of the distribution of global fallout radionuclides (^{3}H, ^{14}C, ^{137}Cs, ^{239,240}Pu) in the ocean basins. ^{3}H and ^{14}C were also used during the most comprehensive tracer program WOCE, which was carried out during the early 1990s [51]. Both programs covered the World Ocean. Some radionuclides, such as ^{137}Cs, ^{90}Sr, and ^{239,240}Pu, were investigated in GEOSECS, International Atomic Energy Agency – Marine Environment Laboratories (IAEA-MEL) and national programs, for example, in the Pacific Ocean [3, 4, 8, 10, 48–50, 52, 53], in the Indian Ocean [54–58], and in the Atlantic and Arctic Oceans [59–61]. Financial support required for large volume water sampling, needed for analysis of ^{137}Cs and ^{239,240}Pu, was a limiting factor for their wider applications in large-scale studies.

One of the most recent and the most comprehensive open ocean studies of anthropogenic radionuclides in the water column has been the IAEA-MEL research program on Worldwide Marine Radioactivity Studies (WOMARS). The primary objective was to develop an understanding of the present open ocean distribution of radionuclides in the water column and sediment, and to contribute to scientific knowledge of the processes, which affect radionuclide behavior in the marine environment. Any contributions over established background levels due to additional nuclear discharges or accidents could then be easily identified [4, 62]. Three anthropogenic radionuclides (^{90}Sr, ^{137}Cs, and ^{239,240}Pu) were chosen as the most important and radiologically typical of each class of marine radioactivity. They are the most abundant anthropogenic radionuclides present in the marine environment, and they may lead to the highest radiation doses to humans and marine biota. The project was thus filling the existing gap in the WOCE program that covered only ^{3}H and ^{14}C. The project was carried out in collaboration with several laboratories in Denmark, France, Germany, India, Italy, Japan, Republic of Korea, New Zealand, Sweden, UK, and USA. New data for ^{3}H, ^{14}C, ^{90}Sr, ^{129}I, ^{137}Cs, ^{238}Pu, ^{239,240}Pu, and ^{241}Am were obtained for the western North Pacific Ocean [10, 63–66] and the adjacent seas [67, 68], for the North and South Indian Ocean [55], including the Arabian Sea [56, 57], and for the Northeastern Atlantic Ocean [69].

This entry reviews the distribution of global fallout radionuclides in World Ocean waters about 5 decades after their main injection from atmospheric nuclear weapons tests on the ocean surface. The entry focuses on waters, because the main idea is to apply radionuclide tracers in studies of ocean currents. The presented results are mainly based on the work carried out in IAEA-MEL and collaborating institutions. Results obtained from the WOMARS project, as well as from the SHOTS (Southern Hemisphere Ocean Tracer Studies) project are compared with some previous measurements. Water samples included in the SHOTS project were collected during the "around-the-globe" BEAGLE2003 expedition (organized by JAMSTEC [70, 71]), which sampled seawater along 20°S–30°S latitude covering the Pacific Ocean [72], the Atlantic Ocean [73], and the Indian Ocean [74]. It was the first time that many radionuclide tracers were used in such a large-scale project producing high-density radionuclide data for water profiles.

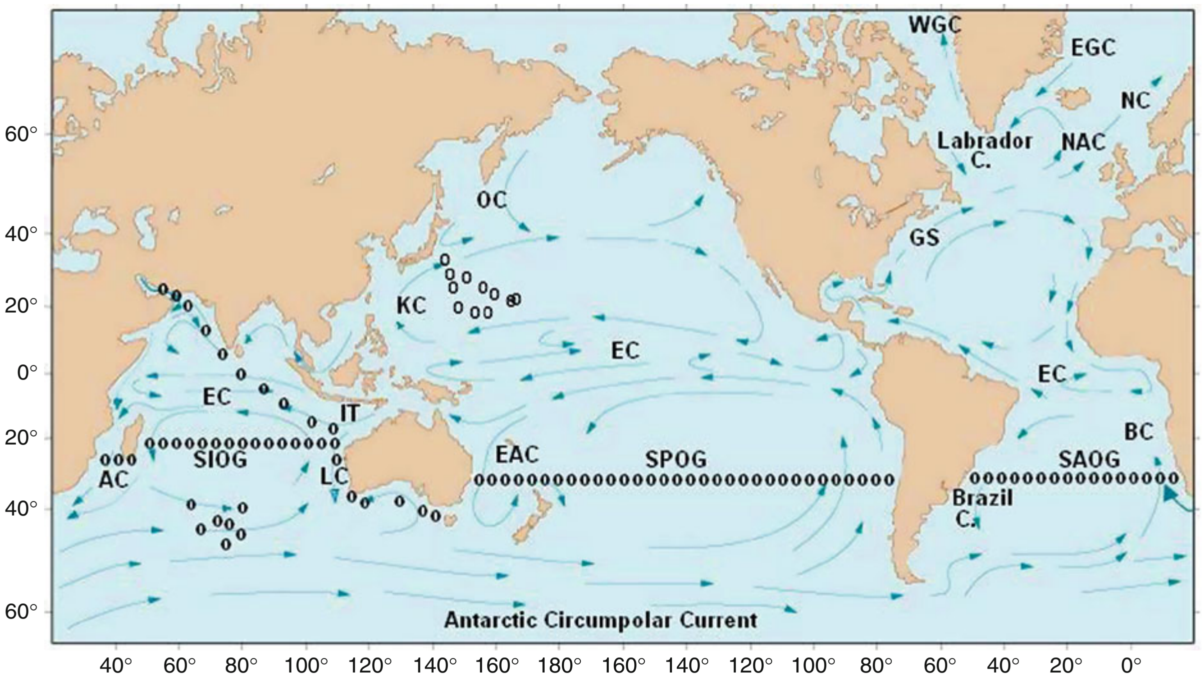
Hydrography Background

Pacific Ocean

The Pacific Ocean is the largest ocean, spreading from the northern boundary of the Bering Sea near 50°N to the boundary of the Southern Ocean at 60°S (Fig. 1). A typical global thermohaline circulation transports cold Circumpolar Bottom Waters (CBW) from the South Pacific to the North Pacific, and upwells deep water to shallower water in the North Pacific Ocean. Warm surface waters in the western North Pacific Ocean move to the Indian Ocean via the Indonesian Seas.

The current system in the western North Pacific Ocean comprises the Kuroshio Current, which is a typical western boundary current consisting of a part of the North Pacific subtropical circulation, and the Oyashio Current, which is included in the North Pacific Subpolar Gyre system. In the eastern North Pacific, the major currents are the Alaska Current included in the Alaskan Gyre, and the California Current, which is a part of the North Pacific Subtropical Gyre. The equatorial current system in the Pacific is complicated as it consists of the North Equatorial Current (NEC), the North Equatorial Countercurrent (NECC), the Northern Subsurface Countercurrent (NSCC), the Equatorial Undercurrent (EUC), the Southern Subsurface Countercurrent (SSCC), and the South Equatorial Current (SEC). Nevertheless, west- and eastward flows are prevailing in the equatorial Pacific Ocean. In the South Pacific Ocean, the South Pacific Subtropical Gyre primarily governs the South Pacific current system, including the East Australian Current (EAC) and the Peru Current.

There is no formation of deep waters in the North Pacific due to freshwater supply, in contrast to the North Atlantic Ocean. However, more complicated subsurface processes, including formation of water masses occur in the North Pacific Ocean. The hydrographic analysis in the North Pacific Ocean suggests that several water masses are formed in the Kuroshio Current and the Kuroshio Extension region. The water mass with high salinity (about ~35) and σ_θ of 23–25, which has been named as the North Pacific Tropical Water (NPTW) [75], is produced in the midlatitude region of the central North Pacific (west of 160°W) as a result of high evaporation and low precipitation. The produced high salinity water is then subducted under the subtropical North Pacific, and transported from the central North Pacific to the western subtropical North Pacific Ocean [76, 77]. The North Pacific Subtropical Mode Water (NPSTMW), which is formed by deep convection in winter, just south of the Kuroshio Current, is defined as a voluminous water mass with an isothermal layer of 16.5°C, σ_θ of approximately 25.5, and a potential vorticity minimum [78]. The North Pacific Central Mode Water (NPCMW), defined as water mass with σ_θ in the range of 26.0–26.4, is formed



Radionuclides as Tracers of Ocean Currents. Figure 1
The main surface currents in the World Ocean with sampling sites visited in the framework of the WOMARS project [62] and the BEAGLE2003 expedition [70, 71] (along 20°S in the Indian Ocean, 32.5°S in the Pacific, and 30°S in the Atlantic Oceans). AC – Agulhas Current; SIOG – South Indian Ocean Gyre; EC – Equatorial Countercurrent; IT – Indonesian Throughflow; LC – Leeuwin Current; KC – Kuroshio Current; Oyashio Current; SPOG – South Pacific Ocean Gyre; EAC – East Australia Current; SAOG – South Atlantic Ocean Gyre; BC – Benguela Current; GS – Gulf Stream; NAC – North Atlantic Current; WGC – West Greenland Current; EGC – East Greenland Current; NC – Norwegian Current (modified from www.AmericanMeteorologicalSociety.org)

around 34°N–41°N and 160°E–165°W [79]. On the other hand, the shallow salinity minimum water (SSM) is present in the eastern North Pacific. The SSM, which may be produced by the subduction of low-salinity surface water in the northern outcropping zone due to Ekman pumping, turns southwestward with the eastern boundary current near 20°N, and spreads between 10°N and 20°N from 130°W to 170° W°W in the density range of 25.4–25.7 $\sigma_\theta\sigma_\theta$ [80, 81]. These water masses, formed in the North Pacific Ocean are closely related to tracer transport in the Pacific Ocean interior.

Indian Ocean

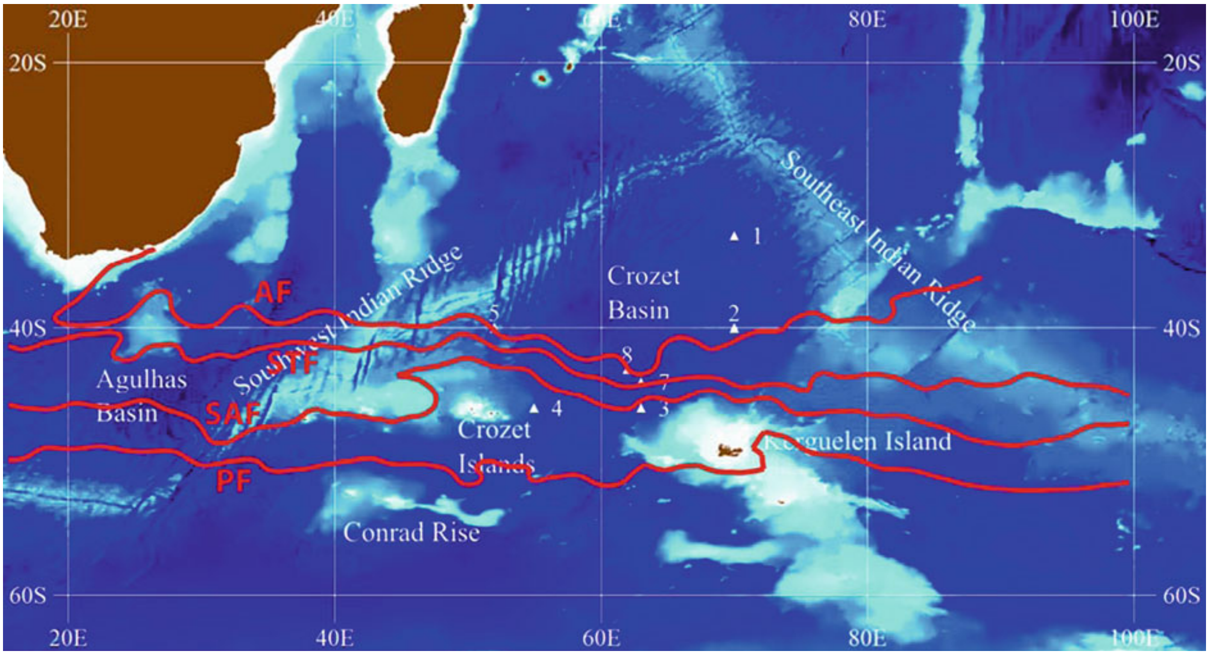
The Indian Ocean is the smallest ocean when compared with the Atlantic and Pacific Oceans as it is limited northward to 25°N (Fig. 1). The southern boundary of the Indian Ocean is set at 60°S, where it becomes the Southern Ocean. A global thermohaline circulation transports warm surface waters from the western North Pacific through the Indonesian throughflow into the Indian Ocean. On the south, the Indian Ocean is via the cold Antarctic Circumpolar Current (ACC) system connected with the Pacific Ocean, and via the Agulhas Current (AC) system with Atlantic Ocean waters [82]. It represents therefore the most fragile part in the global ocean circulation system, important for global climate change studies. As it is surrounded by highly populated continents, there is also a real danger of contamination from land-based sources. The equatorial current system of the Indian Ocean is strongly influenced by the seasonal variation in the winds north of the Equator [82]. From

November to March these winds blow from the northeast (Northeast Monsoon), while from May to September they blow from the southwest (Southwest Monsoon). The change of wind direction north of the Equator then results in a change of currents there. During the Northeast Monsoon season, the North Equatorial Current (NEC) flows westward from 8°N to the Equator; the Equatorial Countercurrent (ECC) flows eastward from the Equator to 8°S and the South Equatorial Current (SEC) flows westward from 8°S to 15°S–20°S. During the Southwest Monsoon the flow north of the Equator is reversed, and a part of the SEC turning north supplies the Somali Current up to the east coast of Africa. The SEC, the Somali Current, and the Monsoon Current then comprise a strong wind-driven gyre in the northern Indian Ocean. Strong upwelling occurs at this time along the Somali and Arabian coasts.

An important feature is a transport of warm water masses (10–15 Sv) from the western North Pacific Ocean via the Indonesian throughflow to the eastern Equatorial Indian Ocean, and their further transport to the South Indian Ocean [83]. The South Indian Ocean Subtropical Gyre is the most important current system influencing water circulation between 20°S and 40°S [82]. Along the eastern boundary of the Indian Ocean, off western Australia, the Leeuwin Current flows poleward along the continental shelf break from about 22°S to 35°S, and then turns eastward. The Leeuwin Current is warm and of relatively low salinity, low dissolved oxygen, and high phosphate content. It transports a significant amount of heat to the south [82].

The southern part of the Indian Ocean (Crozet Basin) is well known for its strong physical currents, which form in this area one of the most dynamic places of the World Ocean. The banded structure of the ACC consists there of narrow jets associated with sharp fronts due to the presence of the Crozet and Kerguelen Islands.

Several frontal systems that meet there, such as Agulhas Front (AF), Subtropical Front (STF), Subantarctic Front (SAF), and Polar Front (PF), represent narrow zones with sharp changes in temperature and salinity [84] (Fig. 2). The dominant current affecting the circulation in the Crozet Basin is the AF,



Radionuclides as Tracers of Ocean Currents. Figure 2
The main currents in the South Indian Ocean (AF – Agulhas Front; STF – Subtropical Front; SAF – Subantarctic Front; PF – Polar Front; modified from [84]), and ANTARES-4 sampling stations [58]

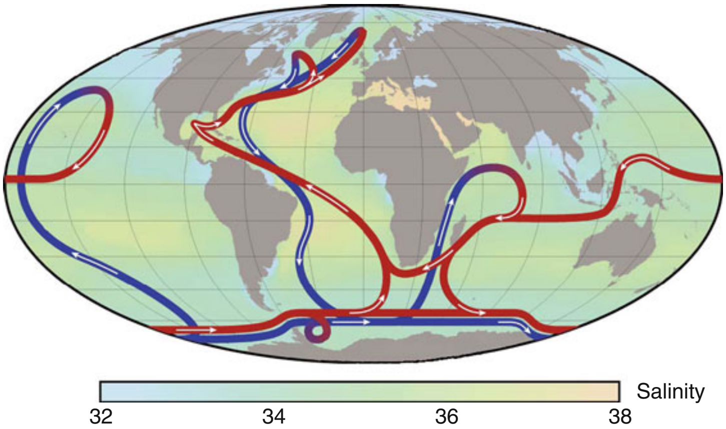
characterized by warm and saline water. Extending eastward into the basin, up to 60°E, AC recirculates to the north, as a part of an anticyclonic Subtropical Gyre [84]. The distribution of tracers in the South Indian Ocean has therefore been controlled by the banded structure of the fronts [85]. The Southern Ocean region has also been documented as a major sink of CO_2, and is thus a highly productive area.

Atlantic Ocean

The Atlantic Ocean is a unique ocean linking with the Arctic Ocean at the north side, with Mediterranean at the east side and with the Weddell Sea in the south side. The main circulation patterns in the Atlantic Ocean are represented by the thermohaline circulation [44] that is part of the global ocean circulation known as the Great Ocean Conveyor Belt (Fig. 3). Warm surface waters from the western South Indian Ocean are advected southward along the eastern African coast by the Agulhas current. A fraction enters the South Atlantic around South Africa forming a branch of the Agulhas Current that does not complete its retroflection, also called the Agulhas leakage [82]. These waters are advected further north and westward, as a part of the Subtropical Gyre of the South Atlantic Ocean, mainly with large eddies or Agulhas rings. A part of these waters continue northward along the western African coast as the Benguela Current, which is responsible for an intergyre water transport to the Central Atlantic Ocean (Fig. 1). The dominant surface water circulation system in the South Atlantic is thus represented by the Subtropical Gyre with the Benguela Current system at the east, and the Brazil Current system at the west (Fig. 1). In the Antarctic Ocean (>60°S), the Weddell Sea Gyre is part of the Polar Current system.

In the Central Atlantic Ocean, the tropical circulation is formed by the Atlantic Equatorial Current system, which consists of the North Equatorial Current (NEC), the North Equatorial Countercurrent (NECC), the Equatorial Undercurrent (EUC), the North Brazil Current (NBC), the Southern Equatorial Countercurrent (SECC), and the South Equatorial Current (SEC). Similarly to the Equatorial Pacific Ocean, west- and eastward zonal flows prevail in the equatorial Atlantic Ocean.

A surface flow in the North Atlantic Ocean, a typical western boundary current, is the Gulf Stream System (Fig. 1), flowing along the Blake Plateau between the Straits of Florida (at ~20°N) and Cape Hatteras (at ~35°N), leaving the continental slope, and after that following the North Atlantic Current as a northeastward branch and the Azores Current as an eastward branch. Much of the water in the North Atlantic Current with the Azores Current water turns southeastward to contribute to the Canary Current. This current system consists of the North Atlantic Subtropical Gyre. In the northern North Atlantic (>50°N), the North Atlantic Subpolar Gyre governs the current system including the Labrador Current and the branch of the North Atlantic Current.



Radionuclides as Tracers of Ocean Currents. Figure 3
Great Ocean Conveyor Belt (modified from www.Wikipedia.org)

The Mediterranean Sea supplies high salinity water (>36) to the North Atlantic Ocean intermediate layer (∼1,000 m), whereas the Arctic Ocean is a source of low-salinity water. Deep water formation that drives the global thermohaline circulation occurs in the Nordic (Greenland, Norwegian and Iceland) Seas and the Labrador and Weddell Seas. It represents therefore the most important part in the global ocean circulation system closely linked with global climate change studies [86]. The North Atlantic Deep Water (NADW) formed in the Nordic and Labrador Seas spreads southward near 50°S, following a flow path near the western boundary. The Antarctic Bottom Water (AABW) formed in the Weddell Sea exits in the region of >30°S. The Antarctic Intermediate Water (AAIW) formed near the Antarctic Polar Frontal Zone intrudes near 1,000 m from 50°S to near the Equator.

Several water masses are formed in the Atlantic Ocean. The North Atlantic Subtropical Mode Water (NASTMW) with properties centered at 18°C, salinity at 36.5 and $26.5\sigma_\theta$ is formed just south of the Gulf Stream Extension. The NASTMW is advected southward out of the Gulf Stream Extension region. The Madeira Mode Water (MMW) with temperature and potential density ranges between 16°C and 18°C and $26.5–26.8\sigma_\theta$ is associated with the warm side of the Azores Front, offshore of the coastal upwelling area. The MMW is advected southwestward from its formation region and joins the thermocline as part of the North Atlantic Central Water. The Subpolar Mode Water (SPMW), which is high-density mode water (density range: $26.9–27.75\sigma_\theta$) in the North Subtropical Gyre and the Subpolar Gyre, originates as thick layers at 14–15°C in the North Atlantic Current loop. The SPMW is advected eastward south of the North Atlantic Current to the eastern Atlantic. The Labrador Sea Water (LSW), which is the fresh intermediate water of the North Atlantic Ocean, arises from convection in the western Labrador Sea. It is advected throughout the subpolar Atlantic, as well as into the subtropical Atlantic.

Sources of Radionuclides in the Marine Environment

Global Fallout Background

The main source of anthropogenic radionuclides in the World Ocean has been global fallout from atmospheric tests of nuclear weapons. A total of 543 nuclear weapons tests have been carried out in the atmosphere with fission yield of 189 Mt of TNT equivalent [2]. The tests with the highest yields were carried out mainly in the late 1950s (1952–1958) and early 1960s (1961–1962). As radionuclides produced in weapons tests with 29 Mt fission yield were not globally dispersed, but were deposited as local fallout (mostly from Bikini and Enewetak atolls); hence only 160 Mt was available for global fallout [2]. The global fallout from the atmospheric nuclear weapons tests has by now been transferred from the atmosphere to the surface of Earth. Since about 70% of the surface of the planet is ocean, much of this mixture of anthropogenic radionuclides has ended up there. Despite the fact that most atmospheric nuclear weapons tests took place in the Northern Hemisphere, it can be shown from the proportions of land and ocean in each hemisphere, and from the latitudinal fallout patterns over time, that over two thirds of the fallout entered the ocean. Following entry to the ocean, the behavior of global fallout radionuclides was determined by their physical and chemical properties and their fate by oceanic physical and biogeochemical processes. As in the terrestrial environment, the postdelivery behavior of these fallout radionuclides provided a unique opportunity to gain insights into the nature of the processes controlling their fate. The great depth and dynamic nature of the oceans have resulted in substantially greater dispersion and dilution than on land.

Initially, the marine radionuclide studies were driven by the need to assess the impact and fate of fallout radionuclides on marine life and associated transfer to man through marine exposure pathways [6]. Subsequently, fallout radionuclides were widely used as oceanic process tracers, due to their unique time-marker labeling of sea water, marine organisms, and sediments according to their individual chemical properties. In this sense, they contributed to the broad growth of knowledge of the intricate physical, chemical, biological, and sedimentological processes, which occur in the oceans. The main periods of delivery of fallout to the oceans were the years following the two intense series of atmospheric tests in the mid-1950s and the early 1960s. The scale of testing by countries that continued atmospheric tests after the 1962 Test Ban Treaty represented a relatively small input. The majority of the fallout produced in atmospheric tests was injected into the stratosphere and then through the specific stratosphere–troposphere mixing returned to Earth's surface as global fallout. Some tests, especially in the earlier test series in the 1950s, were conducted at or near ground level and produced tropospheric fallout with a regional impact. The oceanic impact of ground-level tests was significant from testing at oceanic locations such as those on Pacific islands – especially the U.S. test series at the Pacific Test Site in the Marshall Islands [87]. So this regional impact was largely confined to the western North Pacific. Minor traces of tropospheric fallout from the U.S. Nevada Test Site were identified in the Atlantic Ocean and Mediterranean Sea [88].

The typical depositional pattern with maxima at mid-latitudes and minima toward the Equator and the poles, observed on the land, was also expected on the ocean surface. The patterns of surface seawater concentrations observed during peak fallout years proved rather complex due to seasonal fluctuations in delivery, ocean mixing, and nonstandard observational networks [5, 89]. A convincing demonstration of oceanic distributions came more than a decade later, and relied on water column inventory patterns of ^{137}Cs [90]. The interesting point about these inventory patterns was that not only did they show a latitudinal pattern similar to overland fallout patterns, but that these patterns still were preserved later – despite the dynamic nature of the ocean, with the active gyral circulation of surface and subsurface currents.

The behavior of radionuclides in the marine environment has been classified as either "conservative" or "nonconservative" [90]. The conservative group includes radionuclides having a very small affinity for particles and whose behavior consequently is controlled by the physics of ocean circulation and mixing. Examples of these include ^{3}H, ^{90}Sr, ^{85}Kr, and (in most situations) ^{137}Cs. The nonconservative group includes radionuclides with a range of strengths of particle association and whose ultimate behavior is linked to these affinities. Examples of these include ^{55}Fe, ^{147}Pm, plutonium and americium isotopes, etc. It is important to note that, despite this classification, the timescales over which such behavioral differences may be observed are also important to keep in mind. For example, it is well established that the association of nonconservative radionuclides with plankton blooms provides a mechanism by which reactive fallout nuclides can be removed from surface seawater and released through remineralization processes below the euphotic zone. This can occur over seasonal timescales and produces separation of nonconservative and conservative fallout radionuclides. But as even reactive radionuclides can have oceanic residence times of hundreds to thousands of years, physical circulation processes can dominate in the short term – at least in the open ocean that received the major part of the fallout input.

However, the use of global fallout radionuclides as a tool for ocean science has been limited by the availability and state of development of analytical measurements, sampling technologies, and suitable observation systems. The lack of fixed observational systems covering the oceans (in contrast to the global fallout monitoring on the land), and the enormous area of the oceans, meant that only a very few systematic observations of global fallout delivery to the surface ocean were made. It was noticed [89] that the pattern of seasonal deposition, with a pronounced maximum each spring, was not found generally in over-ocean measurements and that processes of deposition other than precipitation could deliver fallout to the ocean. However, other factors marked the course of the larger-scale studies. These included the way in which ocean science was changing over the years. In the period between 1950 and 1970, ocean studies were carried out largely at the national and institutional level. A new approach to ocean science began later when large-scale projects integrated suites of measurements covering major ocean areas, thus producing in a short time a much better global description of ocean properties or processes. For ocean fallout studies, the GEOSECS project [7–9, 91] was a milestone in marine chemistry and was a major event in the history of ocean fallout studies and their application as tracers of ocean physical and biogeochemical processes. This project spanned the decade of the 1970s with the objective of obtaining a global inventory of chemical components in the world oceans with special emphasis on the deep waters. Intensive sampling was carried out in the Atlantic in 1972–1973, the Pacific in 1973–1974, and the Indian Ocean in 1977–1978. The GEOSECS project produced a comprehensive data set for the fallout radionuclides ^{137}Cs, ^{90}Sr, ^{239,240}Pu (and limited ^{241}Am) for the Atlantic and Pacific Oceans, which permitted inventory estimates for each ocean to be compared with global fallout inputs [8]. Even more detailed data sets on the distributions of ^{3}H and ^{14}C in all three oceans were produced during the GEOSECS project [7, 9].

WOCE was an even larger-scale project, carried out during the early 1990s. It produced the highest data density of ^{3}H [48] and ^{14}C [49] in the World Ocean with important impact not only on the ocean science, but also on general climate change studies [93]. The GEOSECS and WOCE projects led to the specific use of the conservative tracer radionuclides by the physical oceanographic community, and their valuable contributions to ocean science have been widely acknowledged.

The global fallout radionuclides were dispersed horizontally following the main ocean currents, as well as vertically due to specific biogeochemical processes in the water column. During the 1970s, new sources of radionuclides impacted the ocean, and their study and assessment have been made against the preexisting global fallout signal. A comparison of sources of main radionuclide tracers in the World Ocean, represented by cosmogenic radionuclides, global fallout from atmospheric nuclear weapons tests carried out predominantly until 1963, and discharges from nuclear facilities is presented in Table 1.

The distribution of ^{90}Sr, ^{137}Cs, and ^{239,240}Pu in the World Ocean and adjacent seas was treated in the framework of the IAEA WOMARS project [62]. For this purpose, the World Ocean was divided into latitudinal belts for which average radionuclide concentrations were estimated. ^{90}Sr, ^{137}Cs, and ^{239,240}Pu concentrations in surface waters were found to vary considerably, still preserving the latitudinal dependence observed in the 1960s, with lowest values in the southern and highest values in the northern latitudes. The results confirm that the main source of these radionuclides in the marine environment is still global fallout. Time trends in radionuclide concentrations (^{90}Sr, ^{137}Cs, and ^{239,240}Pu) were studied and mean residence times of radionuclides in surface waters of the World Ocean were estimated as well (Table 2). The results confirm a similar residence time for ^{90}Sr and ^{137}Cs in surface waters of 22 ± 2 year. A lower value of 12 ± 1 year has been obtained for ^{239,240}Pu. Average radionuclide concentrations in surface waters estimated for oceans and seas, adjusted to the year 2010, are presented in Table 3. Higher radionuclide concentrations observed in the northeast Atlantic Ocean (the Irish and North Seas) and the Arctic are due to the release and transport of the radionuclides from Sellafield and La Hague reprocessing plants. The Baltic, Black, and Mediterranean Seas were the main reservoirs for radionuclides released after the Chernobyl accident. The ^{90}Sr distribution in surface waters shows a similar pattern to that of ^{137}Cs, confirming the average global fallout ^{137}Cs/^{90}Sr ratio of 1.6 ± 0.1,

Radionuclides as Tracers of Ocean Currents. Table 1 Radionuclide inventories in the World Ocean (after [62])

			Global fallout inventory		Discharges from reprocessing facilities	
Nuclide	Half-life (year)	Natural inventory (PBq)	Total inventory (PBq)	Inventory in 2010 (PBq)	Total inventory (PBq)	Inventory in 2010 (PBq)
^{3}H	12.32	2,200	113,000	8,000	410	45
^{14}C	5,730	1	130	130	2.5	2.5
^{90}Sr	28.78		380	100	7	3
^{137}Cs	30.17		600	170	40	26
^{129}I	15.7×10^{6}	0.6×10^{-3}	0.3×10^{-3}	0.3×10^{-3}	0.04	0.04
^{238}Pu	87.74		0.6	0.4	Regional only	
^{239}Pu	24,110		4	4		
^{240}Pu	6,560		2.5	2.5		
^{241}Pu	14.35		85	8		

with the exception of seas affected by radioactive discharges and the Chernobyl accident.

This entry, however, focuses on those aspects of marine radionuclide studies which contributed to a better understanding of the transport of water masses in ocean basins, as well as transport between the oceans.

Radionuclides as Tracers of Ocean Currents. Table 2 Mean residence time of ^{137}Cs, ^{90}Sr, and ^{239,240}Pu in Pacific, Indian, and Atlantic Ocean surface waters (after [62])

Area	Mean residence time (year)	
	^{137}Cs, ^{90}Sr	^{239,240}Pu
Pacific Ocean	23±4	11±2
Indian Ocean	29±3	13±2
Atlantic Ocean	31±4	13±1
World Ocean	28±2	12±1

Pacific Ocean

Due to specific deposition patterns [5] and the large size of the Pacific Ocean, the largest reservoir of global fallout radionuclides is in this ocean. The depositional situation in the Pacific turned out to be considerably more complex than in the Atlantic. Both the GEOSECS data sets [8], Japanese measurements [94, 95], as well as studies in the USA [96, 97] revealed substantial excesses of measured inventories of ^{137}Cs, ^{90}Sr, and ^{239,240}Pu at many locations in the western North Pacific Ocean compared with the estimated delivery from global fallout. It has become clear that the 68 near-surface nuclear tests [87] conducted at Bikini and Enewetak

Radionuclides as Tracers of Ocean Currents. Table 3 Average radionuclide concentrations in surface waters estimated for oceans and seas, adjusted to the year 2010 (modified after [62])

Sea area	^{90}Sr (mBq L^{-1})	^{137}Cs (mBq L^{-1})	^{239,240}Pu (μBq L^{-1})
North Pacific	0.8±0.1	1.7±0.2	1.2±0.5
Equatorial Pacific	0.9±0.2	1.5±0.2	1.5±0.3
South Pacific	0.6±0.3	0.9±0.4	1.4±0.8
Antarctic	<0.1	<0.1	0.5±0.3
Sea of Japan	0.9±0.2	1.7±0.3	2.4±0.9
Arabian Sea	0.6±0.1	1.0±0.2	0.7±0.4
Indian Ocean	0.8±0.1	1.5±0.2	1.5±0.7
Southern Ocean	0.4±0.2	0.6±0.4	0.2±0.2
Arctic	1.3±0.5	8±4	2.4±0.6
Barents Sea	1.1±0.2	2.2±1.2	7.3±4.4
Baltic Sea	6.2±1.6	37±12	1.3±0.9
North Sea	2.2±0.7	4.1±0.5	5.5±3.7
Irish Sea	27±47	35±33	180±150
English Channel	2.3±0.8	2.6±0.9	4.8±2.9
North N. Atlantic	1.2±0.5	2.6±1.6	13±6
Black Sea	9±3	15±2	1.9±0.8
Mediterranean Sea	0.9±0.1	1.6±0.2	5.2±1.5
North Atlantic	0.7±0.3	1.0±0.5	1.8±1.1
Central Atlantic	0.4±0.1	0.8±0.1	1.0±0.5
South Atlantic	0.2±0.1	0.4±0.1	0.7±0.2

produced not only local fallout but regionally deposited tropospheric fallout, which led to a widespread deposition to the western North Pacific Ocean on a scale similar to that from global fallout. Its presence has been detected not only by the substantial increments to fallout nuclide inventories and concentrations, but also by the distinctive enriched ^{240}Pu/^{239}Pu signature in Pacific seawater and sediments over a wide area [38]. The western North Pacific Ocean is, therefore, unique in the history of atmospheric nuclear weapons test fallout in that it is the one area of the globe where fallout was substantially enhanced over the patterns globally observed.

Contributions from nuclear facilities discharging directly to the Pacific Ocean were negligible when compared with global fallout. They were mostly concentrated on the Japanese coast where a small reprocessing facility operated in Tokai. Recently, a large reprocessing facility has been constructed in Rokkasho area, on the Pacific coast of northern Japan.

A small, but measurable contribution in the ^{137}Cs radioactivity of western North Pacific Ocean surface waters was derived from the Chernobyl accident [25].

Indian Ocean

There were two significant inputs of anthropogenic radionuclides to the Indian Ocean, namely, global fallout and the SNAP-9A satellite RTG burnup [1]. The Indian Ocean because of its smaller size received only 14% of global fallout, in comparison to 52% for the Pacific and 33% for the Atlantic Ocean. Because of the greater spatial coverage of southern latitudes of the Indian Ocean, the global fallout inputs were distributed according to one third part in the northern and two thirds parts in the southern Indian Ocean. Additionally, the Indian Ocean has been receiving global fallout radionuclides by advection of surface waters from the Pacific Ocean via the Indonesian Seas.

The most important contribution of Pu from accidents to global fallout was from the SNAP-9A satellite, which burned up over the Mozambique Channel of the Indian Ocean in 1964 and deposited about 0.6 PBq of ^{238}Pu [98]. This input increased the local surface ^{238}Pu/^{239,240}Pu fallout ratio by a factor of 10 [55].

Discharges from nuclear industry around the Indian Ocean are expected to be small relative to those received by the other oceans. A possible source may be controlled releases from the nuclear facilities at Trombay, near Mumbai (Bombay), which include several research reactors, radioisotope laboratories, and most significantly a nuclear fuel reprocessing plant. The releases peaked in 1971 at 14 TBq of ^{90}Sr equivalent and subsequently decreased below 0.4 TBq [99]. Compared to the releases from Sellafield or Cap de La Hague fuel reprocessing plants, the Trombay inputs seem to have been very small and are probably important only for coastal waters.

Atlantic Ocean

While the Central and South Atlantic Oceans have received inputs of anthropogenic radionuclides only from global fallout, important North Atlantic contributions were made from authorized liquid radioactive discharges from European nuclear fuel reprocessing plants at Sellafield and La Hague, and from the Chernobyl accident. In addition to these sources, nuclear power stations, fuel production facilities, nuclear research facilities, and dumping of low-level radioactive waste in the deep NE Atlantic Ocean [100] may contribute to the contamination of the immediate environments. The discharges or releases from these sources are not detectable in the open marine environment and their impact on the local environment is monitored by the competent authorities [101]. Accidental losses of radionuclide sources on the seabed, including the wreck of the vessel “Carla” on November 30, 1997, close to the Açores with two medical ^{137}Cs sources of 330 TBq and nuclear powered vessels (e.g., Thresher, Komsomolets) carrying nuclear weapons, represent additional potential sources. The nuclear submarine Komsomolets sank on April 7, 1989, in the Norwegian Sea in a depth of 1,680 m with one nuclear reactor and two torpedoes containing about 16 TBq of Pu. However, the limited monitoring which was carried out demonstrated that the contamination was minor and localized [100].

The major source of anthropogenic radionuclides in the NE Atlantic Ocean has been released from two nuclear fuel reprocessing plants. They are the Sellafield plant on the Irish Sea on the northwest coast of England, and the La Hague facility on the northwest coast of France on the English Channel. A third reprocessing plant at Dounreay, on the north coast of the Scottish mainland, also discharges directly into the sea. In general, contamination from this facility has been localized and has tended to have been masked by the Sellafield signal. The first discharge from the Sellafield reprocessing plant (formerly Windscale) took place in 1952 when waste with a total activity of about 370 TBq was discharged to the NE Irish Sea. The overall composition of the discharges has been available since 1960 and more comprehensive data were published by BNFL from 1978 onward and in recent years by the UK Government. This topic has been reviewed in several papers [2, 102]. The discharges of most radionuclides peaked in the mid- to late 1970s, with the peak annual value of 5.2 PBq of ^{137}Cs discharged to the Irish Sea in 1975. A total of 41 PBq of ^{137}Cs were released from 1952 up to the end of 1998. The other long-lived radionuclide discharges totaled about 60 PBq of ^{3}H, 6 PBq of ^{134}Cs, 0.6 PBq of ^{90}Sr, 0.1 PBq of ^{238}Pu, 0.6 PBq of ^{239,240}Pu, 22 PBq of ^{241}Pu, and 1 PBq of ^{241}Am. Most of the plutonium and americium has been retained in the seabed sediments of the Irish Sea. Cesium also became incoporated into sediments and since the mid-1980s, the sediment remobilization has represented a ^{137}Cs source [103].

The La Hague nuclear fuel reprocessing facility has been discharging into the English Channel since 1966 [104]. The discharge levels for most of the radionuclides have been decreasing from a peak value of 243 TBq of ^{137}Cs observed in 1971. Recent discharges are about 2 orders of magnitude lower, with the exception of ^{129}I, which has increased substantially in the past decade. The total ^{137}Cs releases have been estimated at 1 PBq. When compared with Sellafield discharges, the contribution of La Hague to the marine inventory of radionuclides, expressed as a percentage of Sellafield releases, is about 2% for ^{137}Cs, 12% for ^{90}Sr, and 0.4% for ^{239,240}Pu. However, for other radionuclides such as ^{3}H, ^{125}Sb, and ^{129}I, the accumulated discharges from La Hague have been higher by a factor of 2–4. For example, La Hague discharges of ^{129}I amount to about 11 TBq compared to Sellafield's 5 TBq (up to the year 1998). Dispersion of La Hague releases in the English Channel and the North Sea have been reviewed in several papers, and used as tools for the validation of numerical hydrodynamic models [105].

The Chernobyl nuclear accident in April 1986 was the largest nuclear accident to date and has had a significant impact on the marine environment. The total activity of nuclear debris released was high (about 1,100 PBq) and the radioactive fallout was widely distributed after the accident and dominated environmental radionuclide levels in various parts of the world. The total releases of ^{137}Cs were estimated to be about 85 PBq [106]. Of the more than 20 radionuclides that were released in significant quantities during the Chernobyl accident, only a few have been studied in the marine environment. Among the Chernobyl radionuclides, the most important are ^{134}Cs and ^{137}Cs. The cesium isotopes were the most widespread, most abundant, and most investigated of those released. The sea most affected by the Chernobyl accident was the Baltic Sea, as the first radioactive clouds from Chernobyl traveled to the north and caused high deposition over the Scandinavian region. The ^{137}Cs inventory in the Baltic Sea due to the Chernobyl accident increased to about 4.5 PBq [107], higher by at least a factor of 10 than the pre-Chernobyl value. Because of the semi-enclosed nature of this sea, and its limited exchange of water with the North Sea, the levels of ^{137}Cs in the Baltic Sea have remained the highest in Europe, 125 Bq/m^{3} in the reference year of 1990 [108]. The North Sea received about 1.2 PBq of ^{137}Cs by direct deposition. The total Chernobyl inventory in the NES waters was estimated to be about 6 PBq of ^{137}Cs.

In the Arctic Ocean various potential sources exist, such as dumped radioactive wastes in the Kara Sea, discharges from nuclear installations on the Kola Peninsula, and nuclear submarines awaiting decommissioning [15]. Arctic sources include (1) nuclear tests on Novaya Zemlya (NZ), (2) dumping of fully fueled nuclear reactors from submarines and icebreakers in eastern NZ fjords, (3) dumping of liquid wastes in Barents Sea, (4) underwater nuclear tests on NZ, (5) discharges from upstream nuclear facilities on Ob and Yenesey Rivers into Kara Sea, (6) discharges from Murmansk from Russian naval and civilian nuclear facilities, and (7) minor contamination from Kursk accident in Barents Sea.

When comparing the main radionuclide sources in the North Atlantic Ocean it can be seen that the total input due to global fallout was 68 PBq of ^{90}Sr, corresponding to 102 PBq of ^{137}Cs, derived using the

$^{137}Cs/^{90}Sr$ fission yield ratio of 1.5. The estimated values for the region of the North European Seas (NES) are 8 PBq for ^{90}Sr and 12 PBq for ^{137}Cs. The dominant source in NES has been ^{137}Cs from the Sellafield reprocessing plant with about a 68% contribution to the ^{137}Cs inventory [14].

Sampling and Analytical Techniques

This entry reviews radionuclide data obtained from several sampling campaigns carried out in the Pacific, Indian, and Atlantic oceans (Fig. 1). Surface seawater sampling was carried out by pumping 100–200 L of seawater from an average depth of 4 m. Water column samples were taken at different depths from the surface to the bottom either by large volume (200 L) samplers, or by 12 L Niskin bottles. Samples for tritium, radiocarbon, and ^{129}I analyses were stored in 1 L glass bottles with air-tight covering so that no exchange with the surrounding air was possible. Radiocarbon samples were poisoned by adding mercury chloride to prevent any biological activity.

Tritium was measured either by the ^{3}He in-growth method at the University of Miami (USA) or by liquid scintillation spectrometry (after an electrolytic enrichment) at the Institute of Geological and Nuclear Sciences (Lower Hutt, New Zealand) and the IAEA's Isotope Hydrology Laboratory (Vienna, Austria). Tritium results are expressed in Tritium Units (1 TU= 118 $mBq\,L^{-1}$ of water).

Stable isotopes of hydrogen and oxygen (δD, $\delta^{18}O$) were also measured in Lower Hutt and Vienna laboratories. $\delta^{18}O$ analyses were performed using the CO_2-H_2O equilibration procedure. δD analyses were done using the H_2O-Zn reduction method. The isotopic analyses were reported against the international standard VSMOW (Vienna Standard Mean Ocean Water) using conventional delta (δ notation in ‰). The precision of measurements (1σ) was ±0.1‰ for $\delta^{18}O$ and ±1‰ for δD.

Radiocarbon in the form of dissolved inorganic carbon (DIC) was extracted either in the Arizona University AMS facility or in IAEA-MEL [108] by acidification of the water sample to pH ~3. The released carbon dioxide was then (in a flow of high purity oxygen gas) collected in a trap cooled by liquid nitrogen. After purification, the carbon dioxide was finally converted to graphite over a Fe catalyst. The radiocarbon activity in seawater samples is expressed by $\Delta^{14}C$ defined as $\Delta^{14}C = (F_m - 1)10^3$, where F_m (a fraction of modern carbon) is the measured AMS ratio of ^{14}C–^{13}C, normalized to $\Delta^{14}C$ of 25‰.

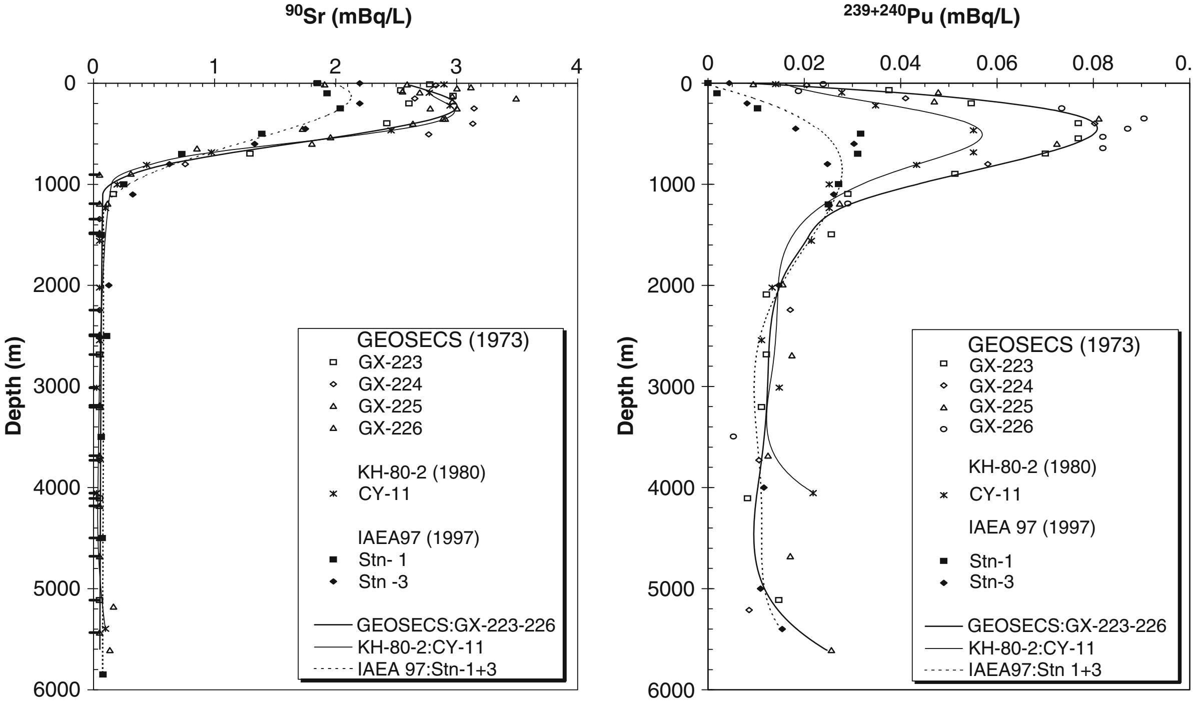
Iodine samples were prepared either at IAEA-MEL or at the IsoTrace Laboratory of the University of Toronto [108]. NaI carrier (10 mg) was added to the sample, and after several steps AgI was obtained. This was used as a target for AMS measurements carried out in the IsoTrace Laboratory. The AMS measurements were normalized with respect to ISOT-2 reference material ($^{129}I/^{127}I=1.174\pm0.022$) 10^{-11}. No background subtraction was required as the machine background was below 2×10^{-14}. The results of ^{129}I analyses of seawater samples are expressed in atom L^{-1}.

Onboard pre-concentration procedures (sequential extraction) were carried out to separate ^{90}Sr, ^{137}Cs, Pu, and ^{241}Am from the collected water samples (MnO_2 co-precipitation). ^{90}Sr was separated by co-precipitation with oxalic acid, and determined in IAEA-MEL by ^{90}Y in-growth method followed by beta-ray counting. Cesium was separated by absorption onto ammonium molybdophosphate and counted in surface or underground gamma-ray spectrometry laboratories. Transuranics were purified using anion exchange resins and extraction chromatography. The samples were then either electrodeposited on stainless steel disks for alpha-ray spectrometry or used for ICPMS or AMS analysis (except ^{241}Am). IAEA reference material, Irish Sea water [109], was analyzed simultaneously with collected samples to ensure high quality of results. A detailed description of the applied procedures used in IAEA-MEL, MRI, and UNIBA is given elsewhere [110–118].

Distribution and Transport of Radionuclides in Seawater – Tracing Ocean Currents

One of the major oceanographic achievements in which radionuclide oceanic tracers played an important role was the development of a model of general circulation between the oceans [44], namely, the Great Ocean Conveyor Belt (Fig. 3). The conveyor belt transfers warm water from the Pacific Ocean to the Atlantic as a shallow current. Passing Europe the surface water evaporates and the ocean water cools, releasing heat to the atmosphere. Colder water with increased salinity becomes dense, and sinks thousands of meters below the surface as it flows from the Nordic Seas into the deep North Atlantic. Cold water from the Atlantic is then transported as a deep current to the Indian Ocean and then to the Pacific Ocean, that flows further to the north up to 1,600 years later. A global ocean circulation between deep, colder water and warmer, surface water strongly influences regional climates around the world. Although this very simplified model was has been further refined, it has helped to understand links between the ocean basins, and the ocean's role in shaping the Earth's climate.

Studies of the evolving three-dimensional distribution of soluble fallout tracers provided a basic set of reference data against which the contrasting behavior of particle active tracers can be compared. For example, the distributions over depth and time of ^{90}Sr and ^{137}Cs have evolved in a tightly coupled way in the open ocean, at least, and can be taken as tracing their dispersion from the ocean surface by physical processes. The behavior of radionuclides in the water column will be further illustrated using radionuclide time series that were developed for a few stations in the Pacific Ocean. By revisiting some of the GEOSECS stations, it was possible to evaluate water column radionuclide data over time. As an example, Fig. 4 compares ^{90}Sr and ^{239,240}Pu profiles for the western North Pacific Ocean as obtained from the GEOSECS expedition [8], from the Hakuho Maru expedition [94], and from the IAEA'97 Pacific Ocean Expedition [10]. Both the ^{90}Sr and ^{239,240}Pu profiles show typical behavior for these radionuclides in the water column. Twenty-four years later, the subsurface ^{239,240}Pu maximum had become much smaller and less pronounced (a decrease by about a factor of 4) and had moved to deeper layers (from 450 to 850 m). The decline in the ^{90}Sr maximum is clearly seen as well (the data were decay corrected to January 1, 1997), although the decrease 24 years later is less than a factor of 2. This observation would emphasize that the observed changes in plutonium over time could not only be caused by the association of



Radionuclides as Tracers of Ocean Currents. Figure 4

Comparison of ^{90}Sr and ^{239,240}Pu profiles as measured by GEOSECS [8], Hakuho Maru [94] and IAEA'97 [10] expeditions in the western North Pacific Ocean (decay corrected to January 1, 1997) plutonium with sinking particles, but also could be due to physical circulation in the upper water column, which has brought water masses bearing significantly lower levels of fallout radionuclides advectively into the region.

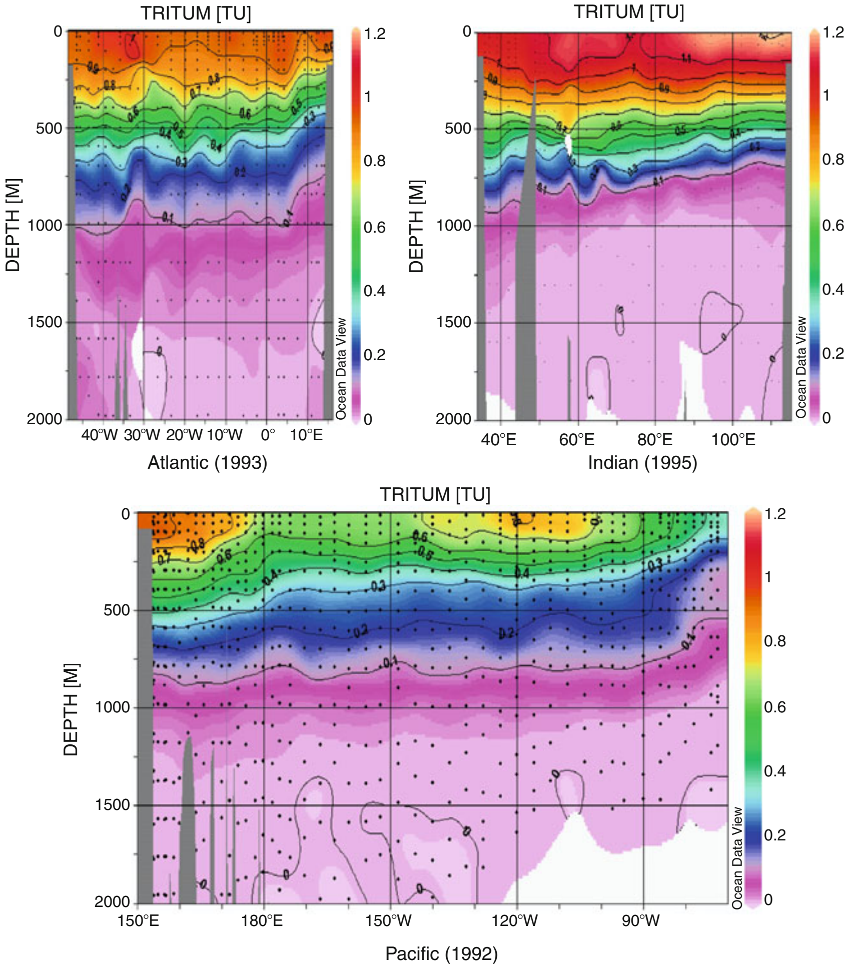
It can be seen from Fig. 4 that the distributions of ^{90}Sr (and similarly for ^{137}Cs) in the water column decline to undetectable levels below 1,000 m, that is, in the main thermocline. In contrast, ^{239,240}Pu continues to be found at depth down to 5,000 m. Due to specific biochemical processes, the major fraction of the ^{239,240}Pu inventory is found therefore in the deep water below the depths containing the major part of the ^{90}Sr and ^{137}Cs inventories. The second broad group of studies was directed specifically to illustrate or quantify specific features of physical mixing [119, 120]. However, the direction of real growth of these studies came in the use of ^{3}H in the large international ocean projects ushered in with the GEOSECS project like Transient Tracers in the Oceans (TTO) in 1981–1983 and South Atlantic Ventilation Experiment in 1988–1989 in the Atlantic Ocean. This could be characterized as the start of recognition by the physical oceanographic community of the potential and power of the use of global fallout ocean tracers in conjunction with standard physical property measurements.

A typical example of the high-density tritium profiles of the World Ocean as obtained from the WOCE program is presented in Fig. 5. The tritium transect was constructed using the WOCE data along 32.5°S in the Pacific Ocean, along 20°S in the Indian Ocean, and along 30°S in the Atlantic Ocean (www.noc.soton.ac.uk/OTHERS/woceipo/atlas_webpage/index.html). In the Pacific Ocean, it is possible to see two tritium maxima: at 155°E (in the Tasman Sea at the western side of the South Pacific Subtropical Gyre), and at 120°W, with penetration depth over 1,000 m. Two ^{3}H maxima ($\sim$1.2 TU) at 100°E and 60°E and penetration of ^{3}H down to 1,000 m water depth can also be seen in the Indian Ocean. The surface maxima coincide with the Subtropical Gyre flow (Fig. 1), maintaining there high ^{3}H levels, comparable with those observed in the western North Pacific Ocean [145]. In the Atlantic Ocean, local, tritium maxima at 5°E and 33°W, which should be associated with the South Atlantic Subtropical Gyre can be seen as well. The tritium penetration depth is over 1,000 m.

In addition to fallout ^{3}H (and its radiogenic progeny ^{3}He), chlorofluorocarbons (CFCs), chemically inert man-made compounds, began to be used in parallel applications as physical tracers. CFCs have been injected into the ocean surface by air–sea gas exchange with an input function that has some similarities to the atmospheric input function of ^{3}H [121]. However, it is noteworthy that the atmospheric history of CFCs, which shows an increase from the 1950s to the end of the 1980s, largely differs from that of ^{3}H. When used together, these tracers have seen the greatest level of adoption for physical tracer use. Both have the advantage that sampling and analysis developed rapidly, permitting the collection of quite high resolution two- and three-dimensional data sets, which found ready assimilation to the physical oceanographic modeling community. As an example of such high-resolution data sets, Schlosser et al. [121] has described the tritium (^{3}He) data sets being developed in the framework of the WOCE program. They give as an example, the north–south line at about 135°W in the Pacific. It is a beautiful snapshot of the near century end distribution of this tracer in the eastern Pacific Ocean and illustrates the application of this fallout tracer in a most impressive way. The WOCE program was strongly connected to ocean/climate studies so the application of the ^{3}H (^{3}He) method as part of this program demonstrates one of the most striking contributions of global fallout to advances in marine science. The ^{3}H (^{3}He) technique has also contributed very specifically and precisely to particular aspects of the study of ocean circulation, namely, the “age” of given water mass. This age, the so-called ^{3}H/^{3}He age, is a measure of time that has elapsed since the last equilibrium of a given water mass with the atmosphere [122]. It rests on the resetting of this ^{3}H/^{3}He clock when a surface mass degasses any helium and the clock begins to record ^{3}He ingrowth when, for example, the body water becomes isolated from the surface by convective cooling and sinking. A classic application of the method was a study on ventilation of water masses in the deep western North Atlantic [123].

Pacific Ocean

Anthropogenic radioactivity monitoring in the 1960s [124] revealed a subsurface maximum of ^{137}Cs



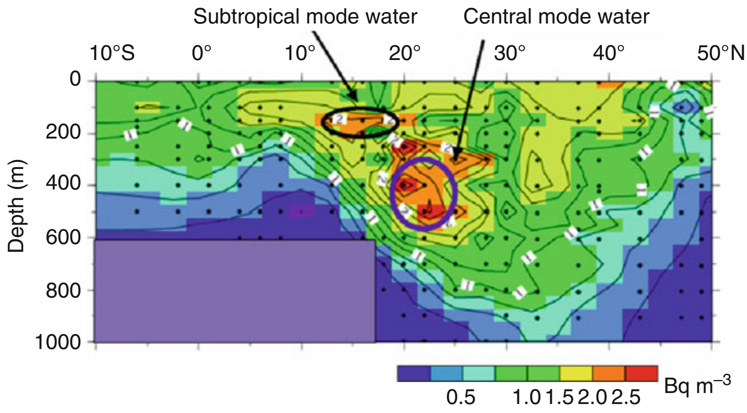
Radionuclides as Tracers of Ocean Currents. Figure 5
Distribution of ^{3}H in the Atlantic Ocean (along 30°S), Indian Ocean (along 20°S), and Pacific Ocean (along 32.5°S) waters (WOCE data: www.noc.soton.ac.uk/OTHERS/woceipo/atlas_webpage/index.html)

(100–200 m depth) in the central and eastern North Pacific water column, which was a first view of subduction in the North Pacific by tracers. The ^{137}Cs and ^{3}H measurements in the Pacific water columns for GEOSECS [8] found that subsurface maxima in the latitudinal vertical sections occurred in the midlatitude and subtropical regions in the North Pacific. A transect of vertical profiles of ^{137}Cs using a composite data set of ^{137}Cs concentrations from the surface to 500-m depth in the western North Pacific (along about 160°W) in the 1970s based on the HAM database [125] indicated that a core water mass with higher ^{137}Cs concentrations exists in a subsurface layer from 10°N to 20°N in the western North Pacific [126], which corresponds to the NPTW with a higher salinity of ∼35. In the eastern North Pacific, the occurrence of a subsurface maximum of bomb ^{3}H in the Subtropical Gyre has been recognized [127]. The water mass with higher ^{3}H seems to correspond to the SSM. A water mass with higher ^{3}H, which was a winter outcrop in the midlatitude of the eastern North Pacific in the early 1960s, penetrated into the eastern subtropical North Pacific. The 1970s ^{137}Cs section in the eastern North Pacific [128] suggested that cores of water masses with higher ^{137}Cs concentrations exist as a subsurface layer (∼100 m depth) from 10°N to 20°N in the eastern subtropical North Pacific, which corresponds to the SSM with lower salinity.

A 1991 ^{3}H section along 135°W (P13C line of WOCE) traces the equatorial transport from the subtropical North Pacific by the ^{3}H tongue, which extends more than 3,000 km meridionally. The ^{3}H tongue extended southward following isopycnals ($26.0 < \sigma_\theta < 26.5$), rising in depth from about 250 m at 30°N to around 75 m at 10°N, and deepening about 100 m at the Equator [129]. A 2002 ^{137}Cs vertical section along 165°E [130] revealed that several cores of ^{137}Cs maxima occurred at 250–500 m depth. A pronounced high ^{137}Cs concentration region, near 500 m depth ($\sigma_\theta \approx 26.0$) at around 20°N marks a major core that corresponds to NPCMW produced in the central North Pacific and a second core near 250 m depth ($\sigma_\theta \approx 25.0$) at 20°N coincides with NPSTMW. Figure 6 shows the schematic diagram of formation, subduction, and transport of NPSTMW and NPCMW.

A Global Ocean Circulation Model (OGCM) can provide information on the subsurface movement of water masses in the North Pacific. The transport of global fallout ^{137}Cs in the Pacific has been studied by Lagrangian trajectory analysis following OGCM streamline [131]. The Lagrangian trajectory analysis suggests that the 1975 subsurface ^{137}Cs maximum water near 10°N along 165°E [126] is identified as the transport of the eastern subtropical Mode Water (EMW), which is formed in the eastern North Pacific around 30°N and 140°W [132], rather than NPSW. The 2002 subsurface ^{137}Cs maximum water near 20°N along 165°E [130] almost coincides with NPCMW, although the Lagrangian particles in 2002 around 20°N spread from the top of the NPCMW-derived water to the bottom of the NPTW-derived water. The tracer experiments reveal that water masses formed in the Kuroshio Current and Kuroshio Extension region are subducted and moved southwestward following the North Pacific subtropical circulation, although the depth of subduction and timescale of subducted water transport depend on the formation region of water mass.

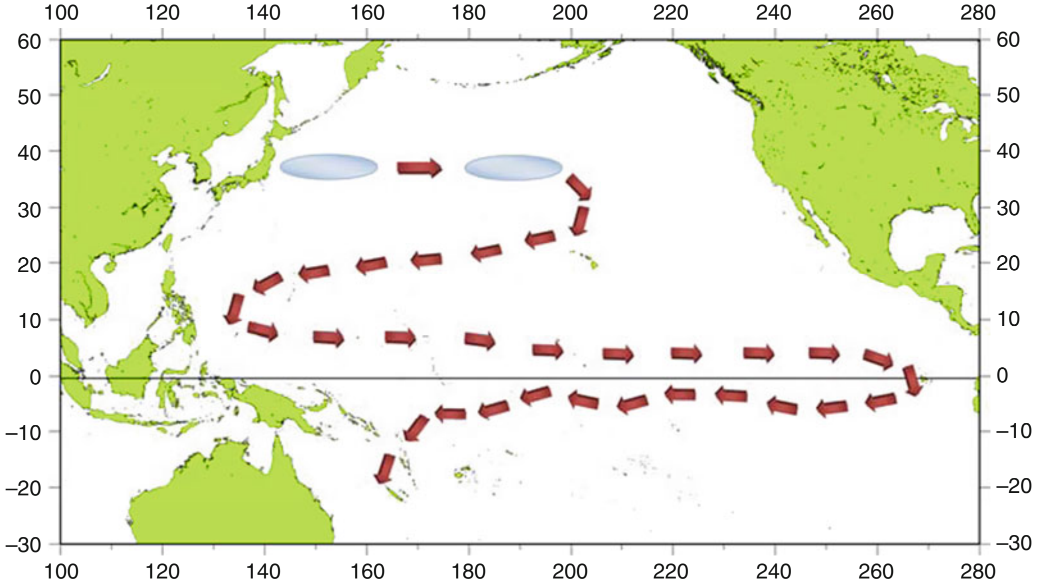
The 1991 ^{3}H and ^{3}He meridional sections along 135°W [129] showed that a strong front exists at the Equator for $24 < \sigma_\theta < 26.5$. This feature is a consequence of the strongly asymmetric global fallout pattern for nuclear explosion–derived radionuclides [125]. There is no obvious indication of cross Equator transport of tracers although ^{3}H isopleths above $\sigma_\theta = 24$ appear to splay, presumably in response to Ekman divergence flow at the Equator. The BEAGLE2003 expedition [72] provided a new view of the ^{137}Cs cross section along 32.5°S, which revealed that an enhanced ^{137}Cs region exists in the Tasman Sea. The higher ^{137}Cs concentration in the Tasman Sea cannot be explained by global fallout and by contamination due to nuclear activities such as French nuclear explosions conducted in the French Polynesia. A possible cause of the enhanced ^{137}Cs in the Tasman Sea is cross Equator transport of ^{137}Cs from the North Pacific to the South Pacific, taken into account temporal changes of surface ^{137}Cs in each sea area of the Pacific,



Radionuclides as Tracers of Ocean Currents. Figure 6
Formation and subduction areas of the North Pacific Subtropical Mode Water (NPSTMW) and the Lighter Central Mode Water (LCMW) (modified from [130])

which suggests Equator ward movement of ^{137}Cs [125]. The Lagrangian trajectory analysis [131] reveals cross Equator pathway from the North Pacific to the South Pacific. The ^{137}Cs in the North Pacific midlatitude region enters the equatorial region through the interior pathways described in previous sections and the Mindanao Current. After that, the ^{137}Cs flows eastward along the EUC and finally enters the eastern South Pacific equatorial region by Ekman transport. A part of ^{137}Cs is transported to the South Pacific Subtropical Gyre via the SEC and East Australian Currents, although ^{137}Cs entering the South Pacific returns to the equatorial region. This pathway is shown in Fig. 7.

Anthropogenic radionuclides and CFCs are strong tools to detect deep water formation, for example, in the Labrador Sea. However, these tracers are insufficient to follow the flow of mid-depth and deep waters in the Pacific because spatial and temporal resolution of concentration contours is too low due to low concentrations to detect water motion. Radionuclide tracers such as ^{3}He and ^{14}C are effective to detect the motion of the mid-depth and deep waters in the Pacific. Submarine hydrothermal fluids that are injected into the ocean interiors from deep-sea ridges and from active seamounts contain enriched ^{3}He comparable to that in the atmosphere. Valuable ^{3}He/^{4}He measurements can be used to follow water movement in mid-depth layers. The larger fraction of excess ^{3}He in the Pacific is injected at the crust of the East Pacific Rise at depths of 2,000–2,500 m, in the eastern equatorial Pacific and eastern South Pacific. A comprehensive excess ^{3}He map in the Pacific was depicted based on WOCE data and others [133]. Two intense ^{3}He-rich jets in the 2,500 m depth surface extend westward into the Pacific basin from the crest of the East Pacific Rise at ≈10°N and ≈15°S with a marked minimum between these plumes centered on the Equator. This feature suggests that westward flow exists at 2,500 m depth in ≈10°N and ≈15°S of the Pacific and eastward flow may be present at the Equator. On the other hand, there is weak indication of southwestward flow in the eastern North Pacific midlatitude region. Another mantle helium injection occurs at hydrothermally active seamounts. Loihi Seamount, on the southeastern flank of the island of Hawaii, has released ^{3}He from active hydrothermal vents centered at 1,100 m depth [134]. The δ^{3}He map at a 1,100-m depth surface reveals that the Loihi ^{3}He extends eastward as a continuous plume



Radionuclides as Tracers of Ocean Currents. Figure 7
Schematic diagram of possible water pathway from the North Pacific midlatitude region (the input region from global fallout) to the Tasman Sea via the Equator

to the coast of Mexico. The helium distribution pattern in the mid-depth of the Pacific is consistent with the circulation pattern predicted by Reid [135] on the basis of a steric height analysis of hydrographic data.

^{14}C is an ideal tracer to study the long-term mean of the deep water flow pattern. The GEOSECS Pacific expedition depicted the meridional Δ^{14}C section in the Pacific, which indicated deep water flow circumpolar water to the North Pacific and allowed us to estimate a net transport of circumpolar water to the North Pacific of 25 Sv [136]. The WOCE Pacific measurements [129] confirmed the GEOSECS data and provided more useful flow pattern data, in which the primary pathway is northward flow as a deep western boundary current passing along the Tonga-Kermadec trench suggested by hydrographic data and the OGCM model. The WOCE Δ^{14}C data can trace the North Pacific deep water flow, which turns westward after across the Equator and then moves clockwise around the basin and finally reaches the Aleutian Island Arch [137]. The bottom water flow in the Pacific is restricted by bottom topography.

In addition to global fallout, there has also been a contribution (especially for plutonium) from close-in fallout from nuclear weapons tests carried out at Bikini and Enewetak Atolls [138]. As many of these tests were conducted at/or near ground level, much of the resultant radionuclides were delivered to the troposphere with substantial fallout to regional areas of the Pacific Ocean, with subsequent transport to the eastern North Indian Ocean. Other Pacific tests conducted at Christmas Island and at Johnson Atoll were carried out at high altitudes, and are unlikely to have contributed significantly to close-in tropospheric fallout in comparison to Enewetak and Bikini tests [3]. A significant amount of close-in fallout Pu was transported in deep layers and dissolved in deep water. As a result, an area with enhanced ^{239,240}Pu derived from Bikini exists in the deep water of the western North Pacific [139]. However, there is no precious map of ^{239,240}Pu in deep layers (>2,000 m depth) of the North Pacific Ocean due to scant observation of deep ^{239,240}Pu. The map of ^{239,240}Pu in deep layers of the North Pacific Ocean would be a valuable tool for tracing deep water flow in the North Pacific Ocean, because deep plutonium behaves as a conservative tracer due to low biological activity in deep waters.

Indian Ocean

Higher radionuclide concentrations were expected to be observed in the areas around 45°N and 30°S–40°S, with values decreasing to the Equator and to the Poles, if the global deposition patterns would have the main impact on radionuclide concentrations [140]. However, from the first comprehensive radionuclide study carried out in the Indian Ocean in 1978 in the framework of the GEOSECS program [9], it was already observed that the latitudinal trend in tritium distribution did not reflect the fallout deposition pattern. Low levels characterized the area south of 40°S, while the highest concentrations were found between 30°S and 10°S. Toward the north, tritium concentration regularly decreased around the Equator, to 10°N and leveled in the Arabian and Red Seas.

The range of observed ^{137}Cs and Pu concentrations in surface waters of the Indian Ocean was between 1.5 and 6 mBq L^{-1} for ^{137}Cs, 0.7–7 μBq L^{-1} for ^{239,240}Pu and 0.6–1.4 μBq L^{-1} for ^{241}Am [54]. These results demonstrated for the first time that a direct transport of ^{137}Cs, ^{239,240}Pu, and ^{241}Am from the Northwest Pacific to the Indian Ocean is feasible via the Indonesian Seas. The data also indicated a radionuclide transport from the eastern Equatorial Indian Ocean via the Leeuwin Current to the eastern South Indian Ocean. A few ^{137}Cs data were obtained during the 1970s [141] (values from 2.8 to 2.9 Bq/m^3) and the early 1990s [142] when the ^{137}Cs concentrations ranged between 1.6 and 2.3 Bq/m^3, while the ^{90}Sr ranged from 1.1 to 1.5 Bq/m^3. Only three results were reported for plutonium isotopes. Again, the lowest levels of ^{137}Cs and ^{239,240}Pu characterized the area south of 30°S. A similar latitudinal trend as obtained for GEOSECS tritium [9] with the smallest values around 60°S and the highest between 10°S and 35°S was observed for the concentration of ^{137}Cs in surface waters. Higher average ^{137}Cs level (2.1 ± 0.3 mBq L^{-1}) was found for the South Indian Ocean, while for the North Indian Ocean, the estimated average concentration is 1.6 ± 0.3 mBq L^{-1}. Low ^{137}Cs concentrations (1.1–1.8 mBq L^{-1}) characterized the whole area south of Australia, at latitudes ranging between 40°S and 35°S. Proceeding northwest, all other stations showed higher ^{137}Cs concentrations (2.1–2.2 mBq L^{-1}). Upwelling near the Equator is probably responsible for the decrease in concentrations in the latitudinal band 10°S and 10°N. ^{239,240}Pu data show a similar latitudinal dependence as do the other radionuclides. The Arabian and Red Seas are characterized by radionuclide concentrations, in the range of 1.4–1.7 Bq/m^3 for ^{137}Cs and 0.9–1.4 for ^{90}Sr. The mean ratio ^{137}Cs/^{90}Sr is 1.63 ± 0.19. As all investigated radionuclides show similar latitudinal trends with the smallest values in the southern box and the highest in the center box between 10°S and 35°S, the broken latitudinal dependence in radionuclide concentrations in surface water (as expected from the global fallout deposition) must be due to specific circulation patterns in the Indian Ocean.

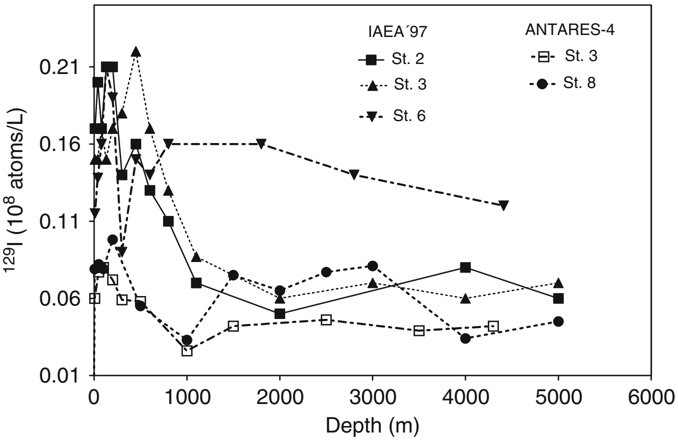
The time-series ^{137}Cs data in the North Pacific show decreasing surface ^{137}Cs concentrations during the past 4 decades, although their decrease rates have been depending on the sea area, being larger in the North Pacific than in the equatorial and South Pacific [10, 143]. As ^{137}Cs after deposition on the ocean surface is affected only by advection, diffusion, and radioactive decay, and there has not been significant transport of ^{137}Cs to bottom waters and to sediments [8, 63, 144], these findings suggest that a significant amount of ^{137}Cs has been transported from the North Pacific Ocean to the Indian Ocean. The observed latitudinal trend, particularly in the Southern Hemisphere, does not reflect that of the integrated deposition density of global fallout radionuclides [3]. Apparently, the system of equatorial currents, the monsoon-induced circulation, transport of water masses from the Mediterranean and from the western North Pacific, outflow to the South Atlantic Ocean, and the Subtropical Gyre have been playing an important role in maintaining radionuclide concentrations in the Indian Ocean surface waters which differ from global fallout.

A comparison of WOMARS tritium data [55, 56] obtained during the late 1990s with GEOSECS (1978) decay corrected values for the Arabian Sea [9] showed unexpectedly high ^{3}H levels at surface and subsurface water depths (0.5–1 TU, i.e., higher almost by a factor of 2). This may be explained by the supply of water masses with higher ^{3}H levels from the Mediterranean Sea, where levels around 2 TU were found [55]. On the eastern side of the Sea there may be also a contribution from the Indonesian throughflow from the western North Pacific Ocean where corresponding ^{3}H levels as high as 1.5 TU were measured [145]. The second interesting feature of this comparison is a deeper penetration of surface ^{3}H into the Indian Ocean water column [56], variable in the depth range of 500–1,500 m. An estimated penetration rate on the basis of ^{3}H and ^{14}C data is around 60 m/year [57]. As there are no water profile data for ^{137}Cs or ^{90}Sr available from the GEOSECS or other cruises, no direct comparison can be made for the Indian Ocean. However, it is possible to compare the ^{137}Cs profiles calculated from the measured GEOSECS ^{3}H profiles using the ^{137}Cs/^{3}H ratio. The observed average ^{137}Cs/^{3}H ratio in the Indian Ocean waters is 0.019 ± 0.004 (the correlation coefficient is 0.74 ± 0.09). This comparison confirms again that higher ^{137}Cs (^{90}Sr) levels were found than expected from the decay corrected ^{3}H GEOSECS data.

Typical profiles of ^{239,240}Pu and ^{241}Am observed in the open ocean have minimum values at the surface, and subsurface maxima at medium water depths. Pu isotopes and ^{241}Am because of their particle-reactive nature are attached to particles, effectively scavenged from surface water, and released again at medium depths by biogeochemical mineralization [7, 10, 146]. The observed Arabian Sea ^{239,240}Pu profiles have maxima at water depths between 400 and 800 m, reflecting different position of sampling sites characterized by slower or faster vertical transport. A similar pattern was observed for the vertical distribution of ^{241}Am, except that the concentrations were lower, and the penetration depths deeper [56]. As Am is more particle reactive than Pu, it reaches deeper depths than Pu because of preferential particulate scavenging and downward transport [143].

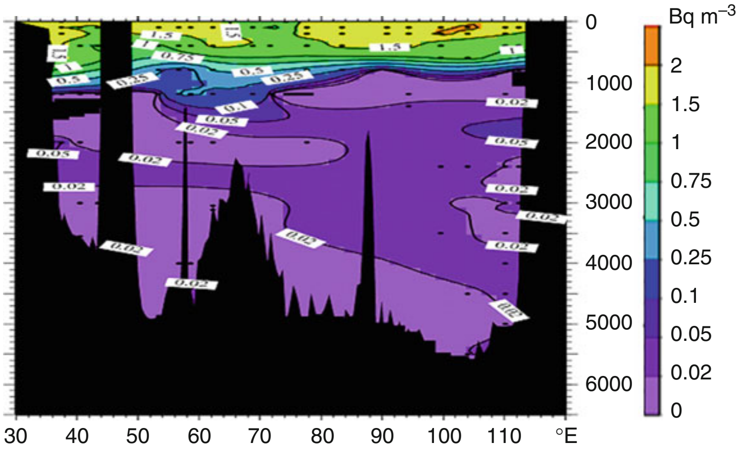
As an example for the South Indian Ocean, ^{129}I water profiles as obtained from the ANTARES-4 cruise will be compared with those measured during the IAEA'97 expedition in the western North Pacific Ocean [145]. Figure 8 documents that similar ^{129}I concentrations in surface seawater were measured during both expeditions. The differences observed in medium depth and deep water samples are due to specific advection conditions observed in the Crozet Basin, where NADW and AABW are manifested, as will be discussed later.

A similar picture on the accumulation of anthropogenic radionuclides in the Subtropical Gyre is presented in Fig. 9. Here, ^{137}Cs profile along 20°S as obtained in the framework of the SHOTS project [74] is showed. This is the first high-density ^{137}Cs profile



Radionuclides as Tracers of Ocean Currents. Figure 8

Comparison of ^{129}I water profile results as obtained during the IAEA'97 expedition in the western North Pacific Ocean with those measured during the ANTARES-4 cruise in the South Indian Ocean (modified from [145])



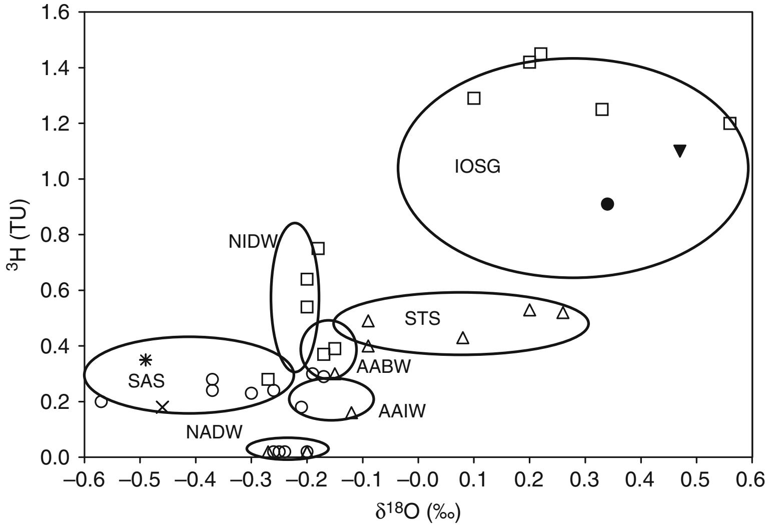
Radionuclides as Tracers of Ocean Currents. Figure 9

Distribution of ^{137}Cs in South Indian Ocean waters along 20°S (data from [74])

obtained thanks to underground gamma-spectrometry analysis of 10–20 L water samples collected during the BEAGLE2003/2004 expedition in the Indian Ocean. The profile shows surface ^{137}Cs maxima at 100°E and 60°E, and deep penetration of ^{137}Cs down to 1,500 m. However, tracers of ^{137}Cs are visible even at depths around 4,000 m. The ^{137}Cs transect is copying the cross section of the gyre (Fig. 1), crossing the main eastern gyre stream roughly at 20°S–100°E. The western gyre, predicted to be at 55°E, has been shifted to 60°E. The central part of the gyre, outside of the main streams, should have the lowest ^{137}Cs concentrations, which is confirmed by observation of lower ^{137}Cs values (1.4 mBq L^{-1}) at 20°S–70°E. This confirms again that transport of water masses from the western North Pacific (where comparable ^{137}Cs were observed 10) occurs via Indonesian Seas to the southern Indian Ocean. The high ^{137}Cs concentration observed in the Madagascar channel (1.95 mBq L^{-1} at 37°31′E) should be due to the Red Sea Overflow Water (Fig. 1), carrying higher ^{137}Cs levels from the Mediterranean Sea [55].

It has been noticed from the ANTARES-4 results (Crozet Basin) that AF, STF, and SAF control radionuclide distribution in surface and subsurface waters with the highest ^{3}H levels observed at the Subtropical Gyre [58, 85]. However, several other water masses can be identified in medium and deep waters [148]. Figure 10 shows the relationship between ^{3}H levels and heavy oxygen isotope content. The intermediate layer (1,000 ±500 m) with decreasing ^{3}H levels represents the Antarctic Intermediate Water (AAIW), which is also characterized by a minimum in ^{129}I at 1,000 m (Fig. 8). The elevated ^{3}H levels around 40°S at water depths 1,500–2,500 may indicate a presence of the North Indian Deep Water (NIDW), which originates from the northern Indian Ocean where comparable ^{3}H levels were observed [55]. The ^{3}H concentrations decrease dramatically below 1,500 m down to 3,000 m, a depth range where the Circumpolar Deep Water (CDW) layer exists. This may be due to the influence of the North Atlantic Deep Water (NADW), which is known to be the most important contributor to the CDW [82]. As measurable ^{3}H concentrations (and negligible levels of ^{129}I) were observed in bottom waters, an injection of Antarctic Bottom Water (AABW) might be considered as the most plausible source of ^{3}H in bottom waters. A major formation area of the AABW is the Weddell Sea, where it is produced by sinking of cold, dense shelf waters. Surface ^{3}H levels in samples collected in 2003 in the Weddell and Ross Seas showed 0.4–0.5 TU, which is consistent with ^{3}H levels found in bottom waters of the Crozet Basin (taking into account a transit time and ^{3}H decay). As the NADW has too low ^{3}H concentrations, and the NIDW is located well above the bottom layer, higher ^{3}H levels found in bottom waters may represent the AABW. This is also confirmed by the fact that the AABW is depleted both in ^{2}H and ^{18}O in comparison with surface waters [85].

The distribution study of radionuclide deposition from nuclear weapons tests clearly showed that the idealized deposition (peaks at 45°N and S) has been interrupted in a few regions, such as western North Pacific Ocean, mainly due to higher precipitation rates [5], impact of close-in fallout from Bikini and Enewetak Atolls, and due to specific seawater circulation patterns. The observed radionuclide levels in the Indian Ocean were significantly influenced by circulation patterns in the North, as well as in the South. Combining the ^{14}C data presented in this work together with WOCE data (www.noc.soton.ac.uk/OTHERS/woceipo/atlas_webpage/index.html), it has been possible to construct a ^{14}C distribution map for



Radionuclides as Tracers of Ocean Currents. Figure 10
Water masses present in the Crozet Basin of the South Indian Ocean (IOSG – Indian Ocean Subtropical Gyre, STS – Subtropical Surface water, SAS – Subantarctic Surface water, NIDW – North Indian Deep Water, NADW – North Atlantic Deep Water, AABW – Antarctic Bottom Water, AAIW – Antarctic Intermediate Water

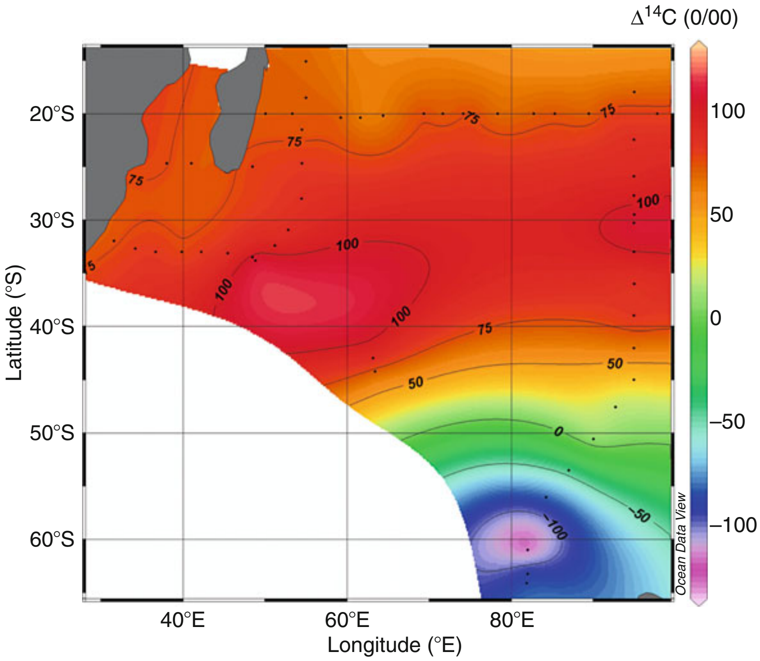
the southern Indian Ocean (Fig. 11). High ^{14}C concentrations can be seen along the Subtropical Gyre (20°S–40°S), with a sharp decrease south of 40°S. This observation has also been confirmed by ^{3}H, ^{90}Sr, and ^{129}I data [85]. When comparing the recent (1997–2003) ^{137}Cs, $^{239,240}Pu$, and ^{241}Am levels with those observed in 1976–1978 [54], it can be seen that there were not big changes in radionuclide concentrations. The ^{137}Cs concentrations observed in surface waters southwest of Australia (influenced by Leeuwin Current) decreased by about a factor of 2 only, in agreement with the estimated mean residence time of ^{137}Cs in the Indian Ocean (29 ± 3 year; Table 2).

A different situation is the case for transuranics, where the observed $^{239,240}Pu$ concentrations are lower by about a factor of 2 than previously reported (the mean residence time of $^{239,240}Pu$ is 13 ± 2 year; Table 2). In the case of ^{241}Am they decreased only by ~30%, in spite of a relatively shorter effective half-life of ^{241}Am (~3 years). A slower decrease in transuranics levels may be due to their additional transport from the Marshall Islands testing grounds, and in the case of ^{241}Am also due to its ingrowth from the decay of ^{241}Pu (half-life 14.4 year).

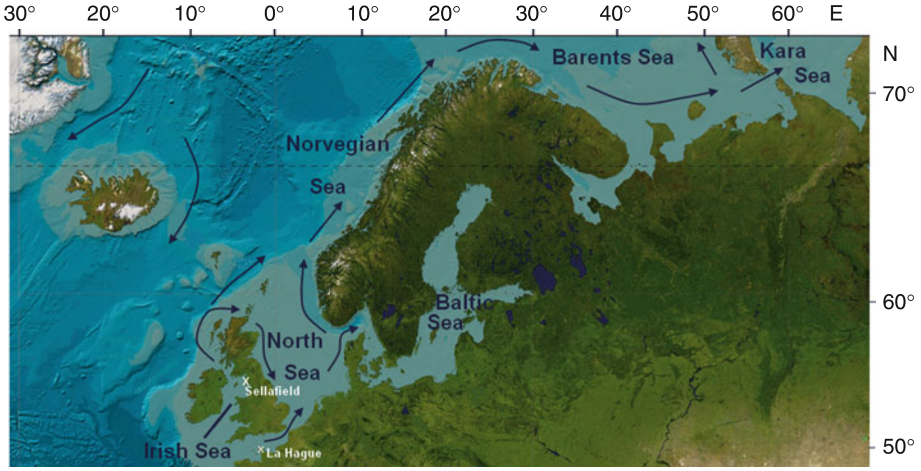
The main feature observed in the South Indian Ocean is the action of the Subtropical Gyre, which concentrates radionuclides between 20°S and 40°S on a timescale of several decades. The Subtropical Gyre due to fresh supply of water masses from the northern latitudes, and their circulation within the Gyre, acts as a reservoir, maintaining higher radionuclide concentrations there.

Atlantic Ocean

The radionuclide tracing of ocean currents in the Atlantic Ocean will be focused on the eastern North Atlantic (Fig. 12), where most of the radionuclide data including tritium, ^{137}Cs, ^{129}I, and ^{99}Tc are available. Although ^{137}Cs is the most frequently analyzed radionuclide in surface waters, the data density has not been high enough for the computer evaluation of data on yearly basis and their presentation in the form of isoline maps. The ^{137}Cs data therefore were grouped in 5-year

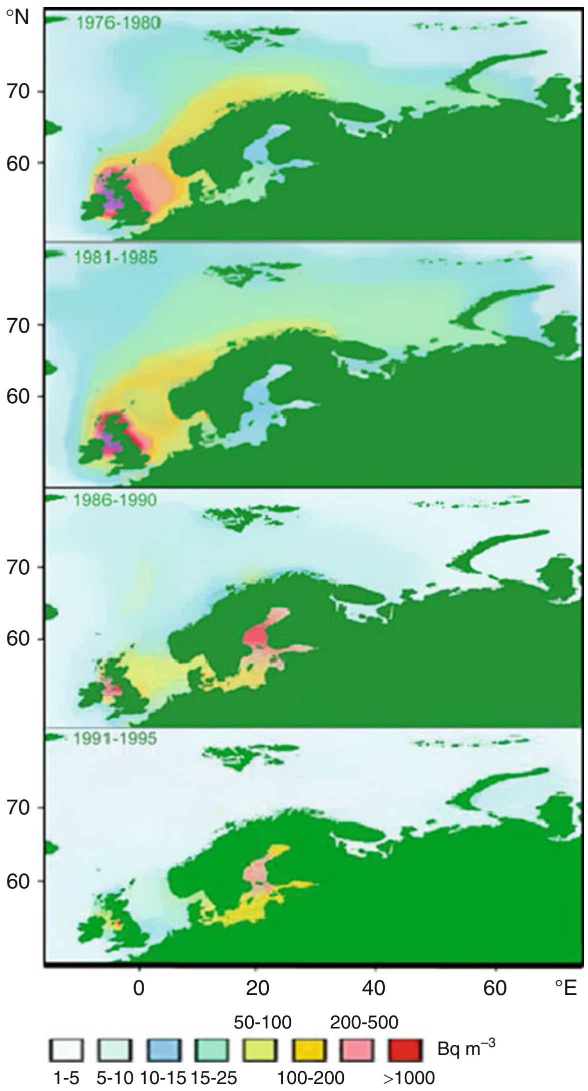


Radionuclides as Tracers of Ocean Currents. Figure 11
Spatial distribution of ^{14}C in surface waters of the South Indian Ocean (combined data of WOCE (www.noc.soton.ac.uk/OTHERS/woceipo/atlas_webpage/index.html) and ANTARES-4 [85])



Radionuclides as Tracers of Ocean Currents. Figure 12
Transport of ^{137}Cs from the Sellafield and La Hague reprocessing nuclear facilities in the eastern North Atlantic Ocean (the map after www.gebco.org)

intervals and evaluated. ^{137}Cs isoline maps computed by data interpolation are presented in Fig. 13, which clearly show the movement of ^{137}Cs contaminated water masses from the Irish Sea through the North Channel to the North Sea, the Norwegian, Barents, and Kara Seas, and onward to the Arctic Ocean. A branch of the North Atlantic Current (NAC) with low (global fallout) ^{137}Cs concentrations entering the northern North Sea is pushing Sellafield contaminated waters to the south, along the east coasts of Scotland and England, before moving eastward, in accordance with the ocean and seafloor topography (Fig. 1). It is because of this specific current system that the ^{137}Cs concentrations along the Scottish, English, and Norwegian coasts are much higher than in the open North and Norwegian Seas. Further, the strong Norwegian Current (NC) keeps the ^{137}Cs concentrations along the coast uniform with high gradients to the west, in contrast to the open Norwegian Sea, where the ^{137}Cs concentrations have been about one order of magnitude lower. This is because of the fact that the open Norwegian Sea is controlled by (NAC), keeping high gradients between ^{137}Cs of NAC and NC waters. The NC receives water from the Baltic and North Seas and from the Norwegian coast. It is a result of wind forcing and freshwater fluxes and it is influenced by bathymetry and the wind-driven cyclonic circulation in the North Sea [149]. Therefore, the transit time necessary for the ^{137}Cs contaminated waters to reach the central Norwegian Sea is much longer. The fast movement of water masses with the NC explains the relatively short transit time to the western South Barents Sea. In the western Barents Sea both NAC and NC warm waters are mixed together, so the west–east ^{137}Cs concentration gradient is lower. The ^{137}Cs lower concentrations observed in the Greenland Sea may be associated with the Greenland Sea Gyre and ventilation of deep waters [150]. At this latitude, a certain portion of the Sellafield-contaminated waters may be transported to the East Greenland, and then transferred back to the Atlantic Ocean through the Denmark Strait [11, 13] (Fig. 12). By comparing the ^{137}Cs release rates from Sellafield and La Hague reprocessing plants with the observed ^{137}Cs distribution, it is possible to derive a transit time for the movement of ^{137}Cs from the Irish Sea to the Nordic Seas (Fig. 12). The transit time of water movement from Sellafield to the North Channel of the Irish Sea can be estimated as 0.5±0.1 year; 2.0±0.2 year to the north of Scotland, 2.5±0.2 year to the east coast of England, 3.0±0.5 year to the southern North Sea, and 4.0±0.5 year to the central and eastern North Sea. In the western Baltic Sea, the signal is detected in 5±1 year after the release into the Irish Sea. Further to the north, peak ^{137}Cs concentrations are found along the Norwegian coast, as well as in the Southwestern Barents Sea, at 4.5 ± 0.5 year after the



Radionuclides as Tracers of Ocean Currents. Figure 13 Distribution of ^{137}Cs from the Sellafield and La Hague reprocessing nuclear facilities (1976–1995) and the Chernobyl accident (1986–1995) in the eastern North Atlantic Ocean and its adjacent seas (modified from [10])

Sellafield discharges. The transit time to the central Barents Sea, as well as to Svalbard is 5 ± 1 year, and 6 ± 1 year to the Kara Sea. The Arctic Ocean was contaminated with ^{137}Cs of Sellafield origin 8 years later. The estimated transit time for ^{137}Cs from the La Hague facility to the North Sea, the Norwegian Sea, etc., would be 2 years less than the transit time estimated from the Sellafield facility. The transit time for water movement south from the Irish Sea to the Celtic Sea is 1.5±0.2 year, and 2.0±0.2 year to the English Channel. The estimated transit times are more precise, and in reasonable agreement with previous estimations [60, 151, 152].

Tritium, ^{129}I, and ^{99}Tc are useful transient tracers as does ^{137}Cs. The GEOSECS 1972 tritium section through the western basin of the North Atlantic [9] highlights the penetration of bomb tritium into the basin over decade timescales, which is the most dramatic evidence for the validity of the "Ocean Conveyer Belt" [44]. This means that recent formed overflow water enters to the Labrador Sea basin from the Greenland/Norwegian Seas and deep convection occurred. ^{129}I and ^{99}Tc have been used to track water flows and NADW formation because discharge of both radionuclides from the Sellafield and La Hague reprocessing plants increased abruptly in the 1990s [153–157]. A marked increase of the ^{129}I concentrations in the Labrador Sea around 2000 reveals that the leading edge of the ^{129}I tracer front derived from the European reprocessing discharge has passed through the Nordic Seas and is entering in NADW via Labrador Sea [158]. The ^{129}I and CFC-11 data in 2001 provide about 2 years for a transit time in Denmark Strait Overflow Water, which is longer than estimates in 1999. This suggests general weakening in the subpolar gyre of the North Atlantic during the 1990s, which may be an indication of oceanic response accompanied with recent climate change.

A specific feature of the water circulation in the North Atlantic is the influence of the Arctic Ocean, which is the smallest, and shallowest of the world's five major oceanic divisions, located mostly in the Arctic north polar region. The Arctic Ocean is partly covered by sea ice throughout the year, and almost completely in winter. Its salinity is the lowest on average of the five major oceans, due to low evaporation, heavy freshwater inflow from rivers, and limited connection with surrounding oceanic waters with higher salinities. The Arctic Ocean contains a major choke point in the southern Chukchi Sea, which provides access to the Pacific Ocean through the Bering Strait between Alaska and Eastern Siberia. The greatest inflow of water comes from the Atlantic via the Norwegian Current, which then flows along the Eurasian coast. Water also enters from the Pacific via the Bering Strait.

The East Greenland Current carries the major outflow. Recently, radionuclide tracers and the water circulation in the North Atlantic and the Arctic Ocean were of interest due to possible transport of radioactive wastes dumped in the Kara Sea [1, 13, 15, 16].

Anthropogenic radionuclide data for the Central and South Atlantic Ocean are very sparse when compared with that in the North Atlantic Ocean. The BEAGLE2003 expedition provided an opportunity to have ^{137}Cs and plutonium data in water columns in the midlatitude region of the South Atlantic Ocean. The ^{137}Cs transect along 30°S shows a core with higher ^{137}Cs concentrations, accompanied with deep intrusion [73]. This ^{137}Cs core corresponds to the Agulhas eddy, which is formed in the Agulhas Current south of South Africa, injected into the Benguela Current and finally carried away northwestward into the South Atlantic Ocean. The ^{137}Cs transect reveals that the Indian Ocean–derived ^{137}Cs trapped in the eddy is transported into the South Atlantic. This is an indication of surface conveyor belt transport from the western North Pacific Ocean to the South Atlantic Ocean via the Indonesian throughflow and the Indian Ocean on a timescale of several decades.

Future Directions

Even 50 years after the first measurements of ^{3}H, ^{14}C, ^{137}Cs, ^{90}Sr, and ^{239,240}Pu were made in the marine environment, there is not at hand a suitable data set (except for ^{3}H and ^{14}C), which could be used for a complex evaluation of radionuclide tracers in the World Ocean. A first look at the global distribution of anthropogenic radionuclides in the water column has been presented in this entry; however, the data density is still too low to allow a more detailed evaluation.

In the Pacific Ocean, tracer (^{3}H, ^{14}C, ^{137}Cs) studies have revealed decadal-scale transport of global fallout radionuclides, which includes ocean interior transport such as subduction and interocean transport. ^{137}Cs has been found to be the most effective tool to trace motion of labeled waters for more than 50 years because of its longer physical half-life and improvement of analytical techniques. New ^{137}Cs data have been accumulated by the Meteorological Research Institute of Tsukuba and by the University of Kanazawa (including data obtained in the framework of the SHOTS project [72]). The ^{137}Cs data set shows new high-resolution latitudinal and longitudinal sections in the Pacific Ocean, in addition to the latitudinal section (at 165°E) and longitudinal sections, obtained from the SHOTS project (at 32.5°S). Another important feature is that the ^{137}Cs data give a constraint to validate the OGCM. The coupling of the OGCM with the ^{137}Cs observations assists in showing the pathway of tracer transport via subsurface currents, including the cross Equator pathway. However, there is still only a vague figure of the tracer transport in deep Pacific waters using tracers due to the low resolution of tracer contours. Another powerful tracer in the Pacific Ocean is ^{239,240}Pu, which has two major sources of global fallout injected in the midlatitude region (30°N–40°N), and close-in fallout in the western subtropical Pacific. Only surface ^{239,240}Pu data have been obtained till now in the SHOTS project for the South Pacific Ocean along 32.5°S latitude [65, 72, 159]. Plutonium behaves as a biogeochemical tracer in shallower layers because of its particle-reactive properties, whereas in deep layers it may move analogously to conservative tracers due to its low biological activity in deep waters. Denser observation of ^{239,240}Pu will be a key factor in solving biogeochemical processes and deep water transport in the Pacific Ocean.

In the Indian Ocean, the ^{137}Cs concentrations observed in surface waters southwest of Australia decreased by about a factor of 2 when compared with 1976–1978 data, in agreement with the estimated mean residence time of ^{137}Cs (29 ± 3 year) in the Indian Ocean. A different situation is in the case of transuranics, where the observed ^{239,240}Pu concentrations are by about a factor of 2 lower than previously reported. In the case of ^{241}Am they decreased only by ~30%, in spite of a relatively shorter mean residence time of ^{241}Am. A slower decrease in transuranics levels may be due to their additional transport from Marshall Islands testing grounds, and in the case of ^{241}Am also due to its ingrowth from the decay of ^{241}Pu. The water column ^{3}H and ^{14}C data obtained during the WOCE program show reasonable density, and new data sets have been obtained by revisiting WOCE stations during the BEAGLE2003 [160] and GEOTRACES expeditions. In the case of ^{137}Cs, only one transect along 20°S latitude was obtained [74]. For ^{239,240}Pu only surface water data are available [159]. It is hoped therefore that more data will be available soon. The dominant circulation system in the South Indian Ocean is represented by the Subtropical Gyre, which accumulates radionuclides on a decadal timescale, as confirmed by ^{3}H and ^{137}Cs observations. This fact is of environmental concern and suggests that the protection of the Indian Ocean against contamination from land-based sources must be considered.

In the Atlantic Ocean, the most comprehensive data set has been obtained for ^{137}Cs in surface waters of the eastern North Atlantic. The dominant actions of the North Atlantic and the Norwegian Currents caused a widespread of ^{137}Cs into the Arctic Ocean. By comparing the ^{137}Cs release rates from the Sellafield and La Hague reprocessing plants with the observed ^{137}Cs distribution, it has been possible to derive a transit time for the movement of ^{137}Cs labeled waters from the Irish Sea to the Arctic Ocean and its adjacent seas. For example, 3 years were required to observe the signal in the southern North Sea, 4 years in the central and eastern North Sea, 5 years in the western Baltic Sea, 6 years in the Kara Sea, and 8 years for the Arctic Ocean. For the South Atlantic Ocean, a first data set on the distribution of ^{137}Cs in surface waters along 30°S latitude obtained in the framework of the SHOTS project [73] has clearly shown transport of western North Pacific waters via Indonesian Seas to the South Indian Ocean, and then via Agulhas Current to the southern Atlantic on a timescale of about 20 years, where they become part of the Benguela Current system and the South Atlantic Subtropical Gyre. For ^{239,240}Pu, only surface water data are available, which also document transport of Pu from the western North Pacific Ocean via Indian Ocean to the South Atlantic Ocean [161].

It is hoped that new developments in radioanalytical techniques based on underground counting systems (e.g., for ^{137}Cs analysis) and on mass spectrometry systems (Accelerator Mass Spectrometry for long-lived radionuclides), which considerably improve the detection sensitivity, and simultaneously permit a smaller sample size, will further stimulate work on radionuclide oceanic tracers. New international projects tracing radionuclides in the World Ocean (e.g., GEOTRACES) will further improve information on oceanic processes, and provide a better understanding of climate change and the protection of the marine environment.

Acknowledgments

Many colleagues contributed significantly to the production of the radionuclide data sets evaluated in the present entry. It is great pleasure to acknowledge long-term collaboration with M. Aoyama (Meteorological Research Institute at Tsukuba), A.J.T. Jull (Arizona University at Tucson), W.E. Kieser (Toronto University), M. Groening (IAEA, Vienna), K. Fiefield (Australian National University of Canberra), U. Morgenstern (Institute of Geological and Nuclear Sciences, Lower Hutt), H. Nies (Bundesamt für Seeschiffahrt und Hydrographie, Hamburg), P. Roos (Risø National Laboratory), B.L.K. Somayajulu (Physical Research Laboratory, Ahmedabad), Z. Top (University of Miami), and former IAEA-MEL directors (M.S. Baxter, H.D. Livingston, R.F.C. Mantoura) and colleagues (J. Gastaud, T. Honda, Y. Ikeuchi, T. Ito, S.-H. Lee, I. Osvath, J. La Rosa, I. Levy, M. Nakano, B. Oregioni, H. Pettersson, O. Togawa, E. Wyse) at Monaco, as well as at the University of Bratislava, and the Meteorological Research Institute of Tsukuba. PPP acknowledges a support provided by the EU Research & Development Operational Program funded by the ERDF (project n. 26240220004).

Bibliography

Primary Literature

1. Livingston HD, Povinec PP (2000) Anthropogenic marine radioactivity. Ocean Coast Manag 43:689–712
2. UNSCEAR (2000) Sources and effects of ionizing radiation. United Nations, New York
3. Livingston HD, Povinec PP (2002) A millennium perspective on the contribution of global fallout radionuclides to ocean science. Health Phys 82:656–668
4. Povinec PP, Hirose K, Honda T, Ito T, Scott ME, Togawa O (2004) Spatial distribution of ^{3}H, ^{90}Sr, ^{137}Cs and ^{239,240}Pu in surface waters of the Pacific and Indian Oceans – GLOMARD database. J Environ Radioact 76:113–137
5. Aoyama M, Hirose K, Igarashi Y (2006) Re-construction and updating our understanding on the global weapons tests ^{137}Cs fallout. J Environ Monit 8:431–438
6. Aarkrog A, Baxter MS, Bettencourt AO, Bojanowski R, Bologa A, Charmasson S, Cunha I, Delfanti R, Duran E, Holm E, Jeffree R, Livingston HD, Mahapanyawong S, Nies H, Osvath I, Pingyu L, Povinec PP, Sanchez A, Smith JN, Swift D (1997) A comparison of doses from ^{137}Cs and ^{210}Po in marine food: a major international study. J Environ Radioact 34:69–90

7. Broecker WS, Peng TH, Östlund G (1986) The distribution of tritium in the Ocean. J Geophys Res 91:14331–14344
8. Bowen VT, Noshkin VE, Livingston HD, Volchok HL (1980) Fallout radionuclides in the Pacific Ocean; vertical and horizontal distributions, largely from GEOSECS stations. Earth Planet Sci Lett 49:411–434
9. Ostlund HG, Brescher R (1982) Tritium laboratory data report no. 2. University of Miami, Miami, USA
10. Povinec PP, Livingston HD, Shima S, Aoyama M, Gastaud J, Goroncy I, Hirose K, Huyhn-Ngoc L, Ikeuchi Y, Ito T, La Rosa J, Liong Wee Kwong L, Lee S-H, Moriya H, Mulsow S, Oregioni B, Pettersson H, Togawa O (2003) IAEA'97 expedition to the NW Pacific Ocean – results of oceanographic and radionuclide investigations of the water column. Deep-Sea Res II 50:2607–2638
11. Nyffeler F, Cigna AA, Dahlgaard H, Livingston HD (1996) Radionuclides in the Atlantic Ocean: a survey. In: Guéguéniat P, Germain P, Metivier H (eds) Radionuclides in the Oceans. Les Editions de Physique, IPSN, Les Ulis, pp 177–197
12. Kershaw PJ, McCubbin D, Leonard KS (1999) Continuing contamination of north Atlantic and Arctic waters by Sellafield radionuclides. Sci Total Environ 237/238:119–132
13. Nies H, Harms IH, Karcher MJ, Dethleff D, Bahe C (1999) Anthropogenic radioactivity in the Arctic Ocean – review of the results from the German project. Sci Total Environ 237/238:181–191
14. Povinec PP, Bailly du Bois P, Kershaw PJ, Nies H, Scotto P (2003) Temporal and spatial trends in the distribution of ^{137}Cs in surface waters of Northern European Seas – a record of 40 years of investigations. Deep-Sea Res II 50:2785–2802
15. Sjoblom K-L, Salo A, Bewers JM, Cooper J, Dyer RS, Lynn NM, Mount ME, Povinec PP, Sazykina TG, Schwarz J, Scott EM, Sivintsev YV, Tanner JE, Warden JM, Woodhead D (1999) International Arctic Seas Assessment Project. Sci Total Environ 237/238:153–166
16. Povinec PP, Osvath I, Baxter MS, Harms I, Huynh-Ngoc L, Liong Wee Kwong L, Pettersson HBL (1997) IAEA-MEL's contribution to the investigation of Kara Sea radioactivity and radiological assessment. Mar Pollut Bull 35:235–241
17. Baxter MS, Harms I, Osvath I, Povinec PP, Scott EM (1997) Modelling the potential radiological consequences of radioactive waste dumping in the Kara Sea. J Environ Radioact 39:161–181
18. Osvath I, Povinec PP, Baxter MS (1999) Kara Sea radioactivity assessment. Sci Total Environ 237/238:167–179
19. Hirose K, Amano H, Baxter MS, Chaykovskaya E, Chumichev VB, Hong GH, Isogai K, Kim CK, Kim SK, Miyao T, Morimoto T, Nikitin A, Oda K, Petterson HBL, Povinec PP, Seto Y, Tkalin A, Togawa O, Veletova NK (1999) Anthropogenic radionuclides in seawater in the East Sea/Japan Sea: results of the first-stage Japanese–Korean–Russian expedition. J Environ Radioact 43: 1–13
20. Ikeuchi Y, Amano H, Aoyama M, Berezhnov VI, Chaykovskaya E, Chumichev VB, Chung CS, Gastaud J, Hirose K, Hong GH, Kim CK, Kim SH, Miyao T, Morimoto T, Nikitin A, Oda K, Pettersson HBL, Povinec PP, Tkalin A, Togawa O, Veletova NK (1999) Anthropogenic radionuclides in seawater of the far Eastern Seas. Sci Total Environ 237/238:203–212
21. Pettersson HBL, Amano H, Berezhnov VI, Chaykovskaya E, Chumichev VB, Chung CS, Gastaud J, Hirose K, Hong GH, Kim CK, Kim SH, Lee SH, Morimoto T, Nikitin A, Oda K, Povinec PP, Suzuki E, Tkalin A, Togawa O, Veletova NK, Volkov Y, Yoshida Y (1999) Anthropogenic radionuclides in sediments in the NW Pacific Ocean and its Marginal Seas. Results of the 1994–1995 Japanese–Korean–Russian expeditions. Sci Total Environ 237/238:213–224
22. Povinec PP, Hirose K, Honda T, Ito T, Marian Scott E, Togawa O (2004) Spatial distribution of ^{3}H, ^{90}Sr, ^{137}Cs and ^{239,240}Pu in surface waters of the Pacific and Indian Oceans – GLOMARD database. J Environ Radioact 76:113–137
23. Povinec PP, Scotto P, Osvath I, Ramadan H (2006) The marine information system. In: Povinec PP, Sanchez-Cabeza JA (eds) Isotopes in environmental studies. IAEA, Vienna, pp 68–69
24. Aoyama M, Hirose K (2004) Artificial radionuclides database in the Pacific Ocean: HAM database. Sci World J 4:200–215
25. Povinec PP, Comanducci JF, Levy-Palomo I (2004) IAEA-MEL's underground counting laboratory in Monaco – background characteristics of HPGe detectors with anti-cosmic shielding. Appl Radiat Isot 61:85–93
26. Povinec PP, Comanducci JF, Levy-Palomo I (2005) IAEA-MEL's underground counting laboratory (CAVE) for the analysis of radionuclides in the environment at very low-levels. J Radioanal Nucl Chem 263:441–445
27. Povinec PP, Comanducci JF, Levy-Palomo I (2006) IAEA-MEL's underground counting laboratory – the design and main characteristics. In: Povinec PP, Sanchez-Cabeza JA (eds) Radionuclides in the environment. Elsevier, Amsterdam, pp 538–553
28. Hirose K, Aoyama M, Igarashi Y, Komura K (2005) Extremely low background measurements of ^{137}Cs in seawater samples using an underground facility (Ogoya). J Radioanal Nucl Chem 263:349–353
29. Hirose K, Aoyama M, Igarashi Y, Komura K (2008) Improvement of ^{137}Cs analysis in small volume seawater samples using the Ogoya underground facility. J Radioanal Nucl Chem 276:795–798
30. Tuniz C, Bird JR, Fink D, Herzog GF (1998) Accelerator mass spectrometry: ultrasensitive analysis for global science. CRC, Boca Raton
31. Kutschera W (2005) Progress in isotope analysis at ultra-trace level by AMS. Int J Mass Spectrom 242:145–160
32. Fifield LK (2008) Accelerator mass spectrometry of long-lived heavy radionuclides. In: Povinec PP (ed) Analysis of environmental radionuclides. Elsevier, Amsterdam, pp 263–295
33. Jull AJT, Burr GS, Beck JW, Hodgins GWL, Biddulph DL, McHargue LR, Lange TE (2008) Accelerator mass spectrometry of long-lived light radionuclides. In: Povinec PP (ed) Analysis of environmental radionuclides. Elsevier, Amsterdam, pp 240–262
34. Lee S-H, Gastaud J, La Rosa JJ, Liong Wee Kwong L, Povinec PP, Wyse E, Fifield LK, Hausladen PA, De Tada LM, Santos GM (2001) Analysis of plutonium isotopes in marine samples by radiometric, ICP-MS and AMS techniques. J Radioanal Nucl Chem 248(3):757–764

35. Wyse EJ, Lee SH, La Rosa J, Povinec PP, de Mora SJ (2001) ICP-sector field mass spectrometry analysis of plutonium isotopes: recognizing and resolving potential interferences. J Anal At Spectrom 16:1107–1111
36. Povinec PP, Betti M, Jull AJT, Vojtyla P (2008) New isotope technologies for environmental physics. Acta Phys Slov 58:1–154
37. Roos P (2008) Analysis of radionuclides using ICP-MS. In: Povinec PP (ed) Analysis of environmental radionuclides. Elsevier, Amsterdam, pp 295–330
38. Buesseler KO (1993) Thermal ionization mass spectrometry. In: Development and evaluation of alternative radioanalytical methods, including mass spectrometry for marine materials. IAEA-TECDOC-683. IAEA, Vienna, pp 45–52
39. Erdmann N, Passler G, Trautmann N, Wendt K (2008) Resonance ionization mass spectrometry for trace analysis of long-lived radionuclides. In: Povinec PP (ed) Analysis of environmental radionuclides. Elsevier, Amsterdam, pp 331–354
40. Weiss W, Roether W (1980) The rates of tritium input into the oceans. Earth Planet Sci Lett 49:435–446
41. Roether W (1994) Studying thermohaline circulation of the ocean by means of tracer data. In: Malanotte-Rizzoli P, Robinson AR (eds) Ocean processes in climate dynamics: global and Mediterranean examples. Kluwer, Dordrecht, pp 157–171
42. Stuiver M, Ostlund HG (1983) GEOSECS Indian Ocean and Mediterranean radiocarbon. Radiocarbon 25:1–29
43. Bard E, Arnold M, Toggweiler JR, Maurice P, Duplessy JC (1989) Bomb ^{14}C in the Indian Ocean measured by accelerator mass spectrometry: oceanographic implications. Radiocarbon 31:510–522
44. Broecker WS (1991) The great ocean conveyor. Oceanography 4:79–89
45. Broecker WS, Bonebakker ER, Rocco GG (1966) The vertical distribution of cesium 137 and strontium 90 in the oceans. J Geophys Res 71:1999–2003
46. Aoyama M, Hirose K (2003) Temporal variation of ^{137}Cs water column inventory in the North Pacific since the 1960s. J Environ Radioact 69:107–117
47. Raisbeck GM, Yiou F (1999) ^{129}I in the oceans: origins and applications. Sci Total Environ 237/238:31–42
48. Livingston HD, Povinec PP, Ito T, Togawa O (2001) The behaviour of plutonium in the Pacific Ocean. In: Kudo A (ed) Plutonium in the environment. Elsevier, Amsterdam, pp 267–292
49. Baxter MS, Fowler SW, Povinec PP (1995) Observations of plutonium in the oceans. Appl Radiat Isot 46:1213–1223
50. Hong GH, Baskaran M, Povinec PP (2004) Artificial radionuclides in the western North Pacific: a review. In: Shiyomi M (ed) Global environmental change in the ocean and on land. Terrapub, Tokyo, pp 147–172
51. Key RM (1996) WOCE Pacific Ocean Radiocarbon Program. Radiocarbon 38:415–423
52. Aoyama M, Hirose K, Miyao T, Igarashi Y, Povinec PP (2001) Temporal variation of ^{137}Cs inventory in the western North Pacific. J Radioanal Nucl Chem 248:785–787
53. Povinec PP, Woodhead D, Blowers P, Bonfield R, Cooper M, Chen QJ, Dahlgaard H, Dovlete C, Fox V, Froehlich K, Gastaud J, Gröning M, Hamilton T, Ikeuchi Y, Kanisch G, Krüger A, Liong Wee Kwong L, Matthews M, Morgenstern U, Mulsow S, Pettersson HBL, Smedley P, Taylor B, Taylor C, Tinker R (1999) Marine radioactivity assessment of Mururoa and Fangataufa Atolls. Sci Total Environ 237/238:249–267
54. Miyake Y, Saruhashi K, Sugimura Y, Kanazawa T, Hirose K (1988) Content of ^{137}Cs, plutonium and americium isotopes in the Southern Indian Ocean waters. Pap Meteorol Geophys 39:95–113
55. Povinec PP, Delfanti R, Gastaud J, La Rosa J, Morgenstern U, Oregioni B, Pham MK, Salvi S, Top Z (2003) Anthropogenic radionuclides in the Indian Ocean surface waters – the Indian Ocean transect 1998. Deep-Sea Res II 50:2751–2760
56. Mulsow S, Povinec PP, Somayajulu BLK, Oregioni B, Liong Wee Kwong L, Gastaud J, Top Z, Morgenstern U (2003) Temporal (3H) and spatial variations of ^{90}Sr, ^{239,240}Pu and ^{241}Am in the Arabian Sea: GEOSECS stations revisited. Deep-Sea Res II 50:2761–2776
57. Bhushan R, Dutta K, Mulsow S, Povinec PP, Somayajulu BLK (2003) Distribution of natural and man-made radionuclides during the reoccupation of GEOSECS stations 413 and 416 in the Arabian Sea: temporal changes. Deep-Sea Res II 50:2777–2784
58. Lee SH, Povinec PP, Gastaud J, Oregioni B, Coppola L, Jeandel C (2009) Radionuclides as tracers of water fronts in the South Indian Ocean – ANTARES IV results. J Oceanogr 65:397–406
59. Holm E, Roos P, Persson RBR, Bojanowski R, Aarkrog A, Nielsen SP, Livingston HD (1992) Radiocaesium and plutonium in Atlantic surface waters from 73°N to 72°S. In: Kershaw PJ, Woodhead DS (eds) Radionuclides in the study of marine processes. Elsevier Applied Science, London, pp 3–11
60. Kershaw PJ, Baxter AJ (1995) The transfer of reprocessing waste from northwest Europe to the Arctic. Deep-Sea Res II 42:1413–1448
61. Livingston HD, Kupferman SL, Bowen VT, Moore RM (1984) Vertical profile of artificial radionuclide concentrations in the central Arctic Ocean. Geochim Cosmochim Acta 48:2195–2203
62. IAEA (2005) Worldwide Marine Radioactivity Studies (WOMARS): radionuclide levels in oceans and seas. IAEA-TECDOC-1439. IAEA, Vienna
63. Moon D-S, Hong G-H, Kim YI, Baskaran M, Chung CS, Kim SH, Lee H-J, Lee S-H, Povinec PP (2003) Accumulation of anthropogenic and natural radionuclides in bottom sediments of the Northwest Pacific Ocean. Deep-Sea Res II 50: 2649–2674
64. Hirose K, Aoyama M, Povinec PP (2003) Concentration of particulate plutonium in surface and deep-water samples collected during the IAEA'97 expedition. Deep-Sea Res II 50: 2639–2647
65. Hirose K, Aoyama M, Fukasawa M, Kim CS, Komura K, Povinec PP, Sanchez-Cabeza JA (2007) Plutonium and ^{137}Cs in surface water of the South Pacific Ocean. Sci Total Environ 381:243–255
66. Lee S-H, Gastaud J, Povinec PP, Hong G-H, Kim S-H, Chung C-S, Lee K-W, Pettersson HBL (2003) Distribution of plutonium and americium in the marginal seas of the northwest Pacific Ocean. Deep-Sea Res II 50:2727–2750

67. Lee S-H, Gastaud J, Povinec PP, Hong G-H, Kim S-H, Chung C-S, Lee K-W, Pettersson HBL (2003) Distribution of plutonium and americium in the marginal seas of the northwest Pacific Ocean. Deep-Sea Res II 50:2727–2750
68. Ito T, Povinec PP, Togawa O, Hirose K (2003) Temporal and spatial variations of anthropogenic radionuclides in Japan Sea waters. Deep-Sea Res II 50:2701–2711
69. Baxter MS, Ballestra S et al (1995) Marine radioactivity studies in the vicinity of sites with potential radionuclide releases. In: Impact of radioactive releases. IAEA, Vienna, pp 125–142
70. Uchida H, Fukasawa M (eds) (2005) WHP P6, A10, I3/I4 Revisit Data Book Blue Earth Global Expedition 2003, vol 1, 2. Aiwa, Tokyo, www.jamstec.org
71. Uchida H, Fukasawa M (eds) (2007) WHP P6, A10, I3/I4 Revisit Data Book Blue Earth Global Expedition 2003, vol 3. Aiwa, Tokyo, www.jamstec.org
72. Aoyama M, Fukasawa M, Hirose K, Hamajima M, Kawano T, Povinec PP, Sanchez-Cabeza J-A. ^{137}Cs in the South Pacific Ocean based on the results of BEAGLE2003 cruises – cross Equator transport of ^{137}Cs from North Pacific Ocean to South Pacific Ocean. Prog Oceanogr, in print for publication
73. Sanchez-Cabeza JA, Levy I, Gastaud J, Eriksson M, Osvath I, Aoyama M, Povinec PP, Komura K. Transport of North Pacific ^{137}Cs labeled waters to the south-eastern Atlantic Ocean. Prog Oceanogr, in print
74. Povinec PP, Aoyama M, Fukusawa M, Hirose K, Komura K, Sanchez-Cabeza J-A, Gastaud J, Ješkovský M, Levy-Palomo I, Sýkora I (2008) Profiles of ^{137}Cs in South Indian Ocean water along the 20°S latitude – an evidence for accumulation of pollutants in the subtropical gyre. Prog Oceanogr, in print
75. Fine RA, Lukas R et al (1994) The equatorial Pacific: a water mass crossroads. J Geophys Res 99:25063–25080
76. Cannon GA (1966) Tropical waters in the western Pacific Ocean, August–September 1957. Deep-Sea Res 13:1139–1148
77. Tsuchiya M (1968) Upper waters of the intertropical Pacific Ocean. Johns Hopkins Oceanographic Study 4:50
78. Suga T, Hanawa K, Toba Y (1989) Sub-tropical mode water in the 137°E section. J Phys Oceanogr 19:1605–1618
79. Suga T, Takei Y, Hanawa K (1997) Thermostat distribution in the North Pacific subtropical gyre: the general mode water and the subtropical mode water. J Phys Oceanogr 27: 140–152
80. Reid JL (1973) The shallow salinity minima of the Pacific Ocean. Deep-Sea Res 20:51–68
81. Yuan X, Talley LD (1992) Shallow salinity minima in the North Pacific. J Phys Oceanogr 22:1302–1316
82. Tomczak M, Godfrey JS (1994) Regional oceanography: an introduction. Pergamon, New York, 422 pp
83. Gordon AL, Susanto RD, Vranes K (2003) Cool Indonesian throughflow as a consequence of restricted surface layer flow. Nature 425:824–828
84. Park YH, Pollard RT, Pollard RT, Read JF, Leboucher V (2002) A quasi-synoptic view of the frontal circulation in the Crozet Basin during the ANTARES-4 cruise. Deep-Sea Res II 49: 1823–1842
85. Povinec PP, Breier R, Coppola L, Groening M, Jeandel C, Jull AJT, Kieser WE, Lee S-H, Liong Wee Kwong L, Morgenstern U, Park Y-H, Top Z (2011) Tracing of water masses using a multi isotope approach in the southern Indian Ocean. Earth Planet Sci Lett, in print
86. Srocker TF, Wright DG (1991) Rapid transitions of the ocean's deep circulation induced by changes in surface water fluxes. Nature 351:729–732
87. Robinson WL, Noshkin VE (1999) Radionuclide characterization and associated dose from long-lived radionuclides in close-in fallout delivered to the marine environment at Bikini and Enewetak Atolls. Sci Total Environ 237/238:311–327
88. Buesseler KO, Sholkovitz ER (1987) The geochemistry of fallout plutonium in the North Atlantic: II. ^{240}Pu/^{239}Pu ratios and their significance. Geochim Cosmochim Acta 51:2623–2637
89. Volchok HL, Bowen VT, Folsom TR, Broecker WS, Schuert EA, Bien GS (1971) Oceanic distributions of radionuclides from nuclear explosions. In: Radioactivity in the marine environment. National Academy, Washington, DC, pp 42–89
90. Livingston HD, Jenkins WJ (1982) Radioactive tracers in the sea. In: The future of oceanography. Springer, New York, pp 163–191
91. Craig H, Turekian KK (1980) The GEOSECS program: 1976–1979. Earth Planet Sci Lett 49:263–265
92. Kupferman SL, Livingston HD, Bowen VT (1979) A mass balance for ^{137}Cs and ^{90}Sr in the North Atlantic Ocean. J Mar Res 37:157–199
93. Schlosser P, Bullister JL, Fine R, Jenkins WJ, Key R, Lupton J, Roether W, Smethie WM (2001) Transportation and age of water masses. In: Siedler GJ, Church J, Gould J (eds) Ocean circulation and climate. Academic, London, pp 431–452
94. Nagaya Y, Nakamura K (1984) ^{239,240}Pu, ^{137}Cs and ^{90}Sr in the central Pacific. J Oceanogr Soc Jpn 40:416–424
95. Nagaya Y, Nakamura K (1987) Artificial radionuclides in the western Northwest Pacific (II): ^{137}Cs and ^{239,240}Pu inventories in water and sediment columns observed from 1980 to 1986. J Oceanogr Soc Jpn 43:345–355
96. Noshkin VE, Bowen VT (1973) Concentrations and distributions of long-lived fallout radionuclides in open ocean sediments. In: Radioactive contamination of the marine environment. IAEA symposium SM-158/45. IAEA, Vienna, pp 671–686
97. Noshkin VE, Robison WL, Wong KM, Eagle RJ (1998) Behavior of plutonium isotopes in the marine environment of Enewetak Atoll. J Radioanal Nucl Chem 234:243
98. National Academy of Sciences (1971) Radioactivity in the marine environment. National Academy of Sciences, Washington, DC. ISBN 0-309-01865-X
99. Patel B, Patel S, Pawar S (1978) Desorption of radioactivity from nearshore sediment. Est Coastal Mar Sci 7:49–58
100. Baxter MS, Ballestra S, Gastaud J, Hamilton TF, Harms I, Huynh-Ngoc L, Osvath I, Parsi P, Pettersson H, Povinec PP, Sanchez A (1996) Marine radioactivity studies in the vicinity of sites with potential radionuclide releases. In: Proceedings of the environmental impact of radionuclide releases. IAEA, Vienna, pp 125–141

101. CEC (1994) The radiological exposure of the population of the European community from radioactivity in the Mediterranean Sea – Project "MARINA-MED". Report MARINA-MED. In: Cigna A (ed) European Commission Report XI-094-93. Brussels
102. Mackenzie AB (2000) Environmental radioactivity: experience from the 20th century – trends and issues for the 21st century. Sci Total Environ 249:313–329
103. Mitchell PI, Condren ON, Leon Vintro L, McMahon CA (1999) Trends in plutonium, americium and radiocaesium accumulation and long-term bioavailability in the western Irish Sea sediment. J Environ Radioact 44:221–250
104. Guéguéniat P, Kershaw P, Herrmann J, Bailly du Bois P (1997) New estimation of La Hague contribution to the artificial radioactivity of Norwegian waters (1992–1995) and Barents Sea (1992–1997). Sci Total Environ 202:249–266
105. Bailly du Bois P, Salomon JC, Guéguéniat P, Gandon R (1995) A quantitative estimate of English Channel water fluxes into the North Sea from 1978 to 1992 based on radiotracer distribution. J Mar Syst 6:457–481
106. WHO (1989) Health hazards from radiocaesium following the Chernobyl nuclear accident. J Environ Radioact 10:257–295
107. HELCOM (1995) Radioactivity in the Baltic Sea 1984–1991. Baltic Sea environment proceedings no. 61. Helsinki Commission, Hamburg. ISSN: 0357-2994
108. Povinec PP, Oregioni B, Jull AJT, Kieser WE, Zhao X-L (2000) AMS measurements of ^{14}C and ^{129}I in seawater around radioactive waste dump sites. Nucl Instrum Meth Phys Res B172:672–678
109. Povinec PP, Badie C, Baeza A, Barci-Funel G, Bergan TD, Bojanowski R, Burnett W, Eikenberg J, Fifield LK, Serradell V, Gastaud J, Goroncy I, Herrmann J, Hotchkis MAC, Ikaheimonen TK, Jakobson E, Kalimbadjan J, La Rosa JJ, Lee S-H, Liong Wee Kwong L, Lueng WM, Nielsen SP, Noureddine A, Pham MK, Rohou J-N, Sanchez-Cabeza A, Suomela J, Suplinska M, Wyse E (2002) Certified reference material for radionuclides in seawater IAEA-381 (Irish Sea Water). J Radioanal Nucl Chem 251:369–374
110. Povinec PP (2004) Developments in analytical technologies for marine radioactivity studies. In: Livingston HD (ed) Marine radioactivity. Elsevier, Amsterdam, pp 237–294
111. La Rosa J, Burnett W, Lee S-H, Levy I, Gastaud J, Povinec PP (2001) Separation of actinides, cesium and strontium from marine samples using extraction chromatography and sorbents. J Radioanal Nucl Chem 248:765–770
112. Lee SH, Gastaud J, La Rosa JJ, Liong Wee Kwong L, Povinec PP (2001) Analysis of plutonium isotopes in marine samples by radiometrics, ICPMS and AMS techniques. J Radioanal Nucl Chem 248:757–764
113. Povinec PP, La Rosa JJ, Lee SH, Mulsow S, Osvath I, Wyse E (2001) Recent developments in radiometric and mass spectrometry methods for marine radioactivity measurements. J Radioanal Nucl Chem 248:713–718
114. Povinec PP (2005) Ultra-sensitive radionuclide spectrometry: radiometrics and mass spectrometry synergy. J Radioanal Nucl Chem 263:413–417
115. Hirose K, Aoyama M, Igarashi Y, Komura K (2005) Extremely low background measurements of ^{137}Cs in seawater samples using an underground facility (Ogoya). J Radioanal Nucl Chem 263:349–353
116. Aoyama M, Hirose K (2008) Radiometric determination of anthropogenic radionuclides in seawater. In: Povinec PP (ed) Analysis of environmental radionuclides. Elsevier, Amsterdam, pp 137–162
117. Povinec PP (2008) Low-level gamma-ray spectrometry for environmental samples. J Radioanal Nucl Chem 276:771–777
118. Levy I, Aoyama M, Hirose K, Povinec PP, Sanchez-Cabeza JA, Azemard S, Eriksson M, Gastaud J, Hamajima Y, Kim CS, Komura K, Osvath I, Roos P, Yim SA. Marine anthropogenic radiotracers in the Southern Hemisphere: new sampling and analytical strategies. Prog Oceanogr, in print
119. Broecker WS (1966) Radioisotopes and the rate of mixing across the main thermocline of the ocean. J Geophys Res 71:5827–5836
120. Rooth CG, Ostlund HG (1972) Penetration of tritium into the Atlantic thermocline. Deep-Sea Res 19:481–492
121. Schlosser P, Bayer R, Boenisch G, Cooper LW, Ekwurzel B, Jenkins WJ, Khatiwala S, Pfirman S, Smethie WM (1999) Pathways and residence times of dissolved pollutants in the ocean derived from transient tracers and stable isotopes. Sci Total Environ 237/238:15–30
122. Jenkins WJ, Clarke WB (1976) The distribution of ^{3}He in the western Atlantic Ocean. Deep-Sea Res 23:481–494
123. Jenkins WJ (1988) The use of anthropogenic tritium and helium-3 to study subtropical gyre ventilation and circulation. Philos Trans R Soc Lond A325:43–61
124. Folsom TR, Grismore R (1970) Survey of Oceanic fallout traces using cesium absorbers. IEEE Trans Nucl Sci 17:202–212
125. Inomata Y, Aoyama M, Hirose K (2009) Analysis of 50-y record of surface ^{137}Cs concentrations in the global ocean using the HAM-global database. J Environ Monit 11:116–125
126. Hirose K, Aoyama M (2003) Analysis of ^{137}Cs and ^{239,240}Pu concentrations in surface waters of the Pacific Ocean. Deep-Sea Res II 50:2675–2700
127. Fine RA, Peterson WH, Ostlund HG (1987) The penetration of tritium into the tropical Pacific. J Phys Oceanogr 17:553–564
128. Hirose K, Aoyama M et al (2006) Plutonium isotopes in seawater of the North Pacific: effect of close-in fallout. In: Povinec PP, Sanchez-Cabeza JA (eds) Radionuclides in the environment, pp 67–82
129. Schlosser P, Bullister JL et al (2001) Transformation and age of water masses. In: Siedler G et al (eds) Ocean circulation & climate, observing and modelling the global ocean, vol 77, International geophysics series. Academic, New York, pp 431–452
130. Aoyama M, Hirose K et al (2008) Water masses labeled with global fallout ^{137}Cs formed by subduction in the North Pacific. Geophys Res Lett 35:L01604
131. Nakano H, Motoi T et al (2010) Analysis of ^{137}Cs concentration in the Pacific using a Lagrangian approach. J Geophys Res (in press)

132. Hautala S, Roemmich DH (1998) Subtropical mode water in Northeast Pacific Basin. J Geophys Res 103:13055–13066
133. Lupton JE (1998) Hydrothermal helium plumes in the Pacific Ocean. J Geophys Res 103:15853–15868
134. Lupton JE (1996) A far-field helium plume from Loihi Seamount. Science 272:976–979
135. Reid JL (1997) On the total geotrophic circulation of the Pacific: flow patterns, tracers, and transports. Prog Oceanogr 39:263–352
136. Stuiver M, Quay PD, Ostlund HG (1983) Abyssal water carbon-14 distribution and the age of the world oceans. Science 219:849–851
137. Key RM, Kozyr A, Sabine CL, Lee K, Wanninkhof R, Bullister JL, Feely RA, Millero FJ, Mordy C, Peng T-H (2004) A global ocean carbon climatology: Results from Global Data Analysis Project (GLODAP), Global Biogeochem Cycles, 18:GB4031. doi:10.1029/2004GB002247
138. Mittelstaedt E, Osvath I, Povinec PP, Togawa O, Scott EM (1999) Transport of radionuclides from the Mururoa and Fangataufa Atolls through the marine environment. Sci Total Environ 237/238:301–309
139. Hirose K, Aoyama M, Kim CS, Kim CK, Povinec PP (2006) Plutonium isotopes in seawater of the North Pacific: effect of close-in fallout. In: Povinec PP, Sanchez-Cabeza JA (eds) Radionuclides in the environment, IAEA, Vienna, pp 67–82
140. Aarkrog A (2003) Input of anthropogenic radionuclides into the World Ocean. Deep-Sea Res II 50:2597–2606
141. Holm E, Roos P, Josefsson D, Persson B (1996) Radioactivity from the North Pole to the Antarctic. In: Guéguégnat P, Germain P, Métivier H (eds) Radionuclides in the oceans, inputs and inventories. Les Editions de Physique, Les Ulis, pp 59–74
142. Bourlat Y, Millies-Lacroix J-C, Le Petit G, Bourguignon J (1996) ^{90}Sr, ^{137}Cs and $^{239+240}$Pu in world ocean water samples collected from 1992 to 1994. In: Guéguégnat P, Germain P, Métivier H (eds) Radionuclides in the oceans, inputs and inventories. Les Editions de Physique, Les Ulis, pp 75–93
143. Livingston HD, Povinec PP, Ito T, Togawa O (2001) The behaviour of plutonium in the Pacific Ocean. In: Kudo A (ed) Pu in the environment. Elsevier, Amsterdam, pp 267–292
144. Lee S-H, Povinec PP, Wyse E, Pham MK, Hong G-H, Chung CS, Kim S-H, Lee HJ (2005) Distribution and inventories of ^{90}Sr, ^{137}Cs, ^{241}Am and Pu isotopes in sediments of the Northwest Pacific Ocean. Mar Geol 216:249–263
145. Povinec PP, Lee SH, Liong Wee Kwong L, Oregioni B, Jull AJT, Kieser WE, Morgenstern U, Top Z (2010) Tritium, radiocarbon, ^{90}Sr and ^{129}I in the Pacific and Indian Oceans. Nucl Instrum Meth Phys Res B268:1214–1218
146. Yamada M, Zheng J, Wang Z-L (2006) ^{137}Cs, $^{239+240}$Pu and ^{240}Pu/^{239}Pu atom ratios in the surface waters of the western North Pacific Ocean, eastern Indian Ocean and their adjacent seas. Sci Total Environ 366:242–252
147. Hirose K, Aoyama M, Povinec PP (2009) ^{239,240}Pu/^{137}Cs ratios in the water column of the North Pacific: a proxy of biochemical processes. J Environ Radioact 100:258–262
148. Coppola L, Roy-Barman M, Mulsow S, Povinec PP, Jeandel C (2005) Low particulate organic carbon export in the frontal zone of the Southern Ocean (Indian sector) revealed by ^{234}Th. Deep-Sea Res I 52:51–68
149. Mork M (1981) Circulation phenomena and frontal dynamics of the Norwegian coastal current. Philos Trans R Soc Lond Ser A 302:635–647
150. Livingston HD (1988) The use of Cs and Sr isotopes as tracers in the Arctic Mediterranean Seas. Philos Trans R Soc Lond Ser A 325:161–176
151. Kautsky H (1998) Determination of distribution processes, transport routes and transport times in the North Sea and the Northern North Atlantic using artificial radionuclides as tracers. In: Guary JC, Guéguéniat P, Pentreath RJ (eds) Radionuclides: a tool for oceanography. Elsevier, Amsterdam, pp 271–280
152. Dahlgaard H (1993) Where does all the Cs and Sr in Greenland waters come from? In: Proceedings of the environmental radioactivity in the Arctic and Antarctic. NRPA, Osteras, pp 121–124
153. Edmonds HN, Smith JN, Kilius LR, Livingston HD, Edmond JM (1998) ^{129}I in archived seawater samples: source functions and tracer comparisons. Deep-Sea Res I 45:1111–1125
154. Brown JE, Isojpe M, Kolstad AK, Lind B, Rudjord AL, Stand P (2002) Temporal trends for ^{99}Tc in Norwegian coastal environments and spatial distribution in the Barents Sea. J Environ Radioact 60:49–60
155. Karcher MJ, Gerland S, Harms I, Isojpe M, Heldal H, Kershaw PJ, Sickel M (2004) The dispersion of technecium-99 in the Nordic Seas and the Arctic Ocean: a comparison of model results and observation. J Environ Radioact 74:185–198
156. Smith JN, Ellis KM, Kilius LR (1998) ^{129}I and ^{137}Cs tracer measurements in the Arctic Ocean. Deep-Sea Res I 45(6):959–984
157. Smith JN, Ellis KM, Boyd TM (1999) Circulation features in the Central Arctic Ocean revealed by nuclear fuel reprocessing tracers from SCICEX 95 and 96. J Geophys Res 104(C12): 29,663–29,677
158. Smith JN, Jones EP, Moran SB, Smethie WM Jr, Kieser WE (2005) ^{129}I/CFC-11 transit times for Denmark Strait overflow water in the Labrador and Irminger Seas. J Geophys Res 110: C05006. doi:10.1029/2004JC002516
159. Hirose K, Kim CS, Yim SA, Aoyama M, Fukasawa M, Komura K, Povinec PP, Sanchez-Cabeza J-A. Vertical profiles of plutonium in the central South Pacific. Prog Oceanogr, in print
160. Kumamoto Y, Murata A, Watanabe S, Fukasawa M. Temporal and spatial variations in bomb-produced radiocarbon along BEAGLE2003 lines – revisits of WHP P06, A10 and I03/I04 in the Southern Hemisphere oceans. Prog Oceanogr, in print
161. Gastaud J, Povinec PP, Aoyama M, Hirose K, Sanchez-Cabeza JA, Levy I, Roos P, Eriksson M, Bosc E. Long-term scavenging of Pu from surface waters of the Southern Hemisphere Oceans. Prog Oceanogr, in print

Books and Reviews

Aoyama M, Povinec P, Sanchez-Cabeza JA (eds) (2010) South Hemisphere ocean tracer studies. Prog Oceanogr, submitted (special issue)

Broecker WS, Peng WH (1982) Tracers in the sea. Eldigo, New York, 690 pp

IAEA (2005) Worldwide Marine Radioactivity Studies (WOMARS): radionuclide levels in oceans and seas. IAEA-TECDOC-1439. IAEA, Vienna

Livingston HD (ed) (2004) Marine radioactivity. Elsevier, Amsterdam, 301 pp

Povinec P (ed) (1986) Low-level counting. Proceedings of low radioactivities ’85. Nucl Instrum Meth Phys Res B17:377–588

Povinec P (ed) (1987) Low-level counting and spectrometry. VEDA, Bratislava, 308 pp

Povinec P (ed) (1991) Rare nuclear decays and fundamental physics. J Phys G 17(special issue):543

Povinec P (1992) Rare nuclear processes. World Scientific, Singapore, 441 pp

Povinec PP (ed) (1999) Marine environment – understanding and protecting for the future. Sci Total Environ 237/238(special issue):1–526

Povinec PP (ed) (2008) Analysis of environmental radionuclides. Elsevier, Amsterdam

Povinec PP, Sanchez-Cabeza JA (eds) (2006) Isotopes in environmental studies. IAEA, Vienna, 710 pp

Povinec PP, Sanchez-Cabeza JA (eds) (2006) Radionuclides in the environment. Elsevier, IAEA, Vienna, 646 pp

Povinec PP, Betti M, Jull AJT, Vojtyla P (2008) New isotope technologies for environmental physics. Acta Phys Slov 58:1–154

Rahmstorf S (2006) Thermohaline ocean circulation. In: Elias SA (ed) Encyclopedia of quaternary sciences. Elsevier, Amsterdam

Tomczak M, Godfrey JS (1994) Regional oceanography: an introduction. Pergamon, New York, 422 pp

Tuniz C, Bird JR, Fink D, Herzog GF (1998) Accelerator mass spectrometry: ultrasensitive analysis for global science. CRC, Boca Raton, 371 pp

UNSCEAR (1993) Sources and effects of ionizing radiation. United Nations, New York, 922 pp

UNSCEAR (2000) Sources and effects of ionizing radiation. United Nations, New York

Rainwater Harvesting

Sangho Lee, Reeho Kim
Department of Water Resources and Environmental Research, Korea Institute of Construction Technology, Goyang-Si, Gyeonggi-Do, Republic of Korea

Article Outline

Glossary
Definition of the Subject
Introduction
Describing Rainwater Harvesting
Technologies for Rainwater Harvesting
Future Directions
Bibliography

Glossary

Catchment A surface that collects rainwater, which is to be used in a beneficial manner.

Cistern Storage tank.

Detention Detention systems provide a temporary storage of runoff for subsequent release.

First flush Initial surface runoff of a rainfall.

First flush effect Rapid changes in water quality (pollutant concentration or load) that occur in the initial stage of a rainfall.

Health risk assessment An evaluation of the potential for adverse health effects to occur as a result of actual or potential exposures to chemicals.

Hydrological cycle Hydrological cycle (also know as Hydrologic cycle or water cycle) Continuous movement of water on, above, and below the surface of the Earth.

Irrigation Supply of land or crops with water.

Nonpotable water Water intended for uses other than potable purposes.

Nonpoint source pollution Pollution whose sources cannot be pinpointed but rather is washed from the land surface in a diffuse manner by stormwater runoff.

Pathogens Disease-causing organisms capable of inflicting damage on a host it infects.

Rainwater harvesting Direct collection of precipitation falling on the roof or onto the ground without passing through the stage of surface runoff on land. The collected rainwater is mostly used for non-potable but may be used for potable use after proper treatment steps.

Reclaimed water Municipal wastewater that has gone through various treatment processes to meet specific water quality criteria with the intent of being used in a beneficial manner.

Potable water Water deemed safe for human consumption, food preparation, and bathing.

Rainwater Water that has fallen as rain

Rainwater storage tank A tank that stores collected rainwater prior to its beneficial use.

Retention Retention systems retain stormwater without subsequent surface discharge.

Rooftop The outer surface of a roof.

Stormwater Precipitation that is discharged across the land surface or through conveyances to one or more waterways and that may include storm water runoff, snow melt runoff, and surface runoff and drainage.

Surface runoff Water flow that occurs when soil is infiltrated to full capacity and excess water from rain, snowmelt, or other sources flows over the land.

Sustainability The principle of optimizing the benefits of a present without compromising the ability of future generations to meet their own needs.

Trace organics Organic compounds detected at very low (minute) levels by the use of sophisticated instrumentation capable of measuring concentrations in the range of 10^{-12}–10^{-3} mg/L.

Wastewater Used water discharged from homes, business, cities, industry, and agriculture. Various synonymous uses such as municipal wastewater (sewage), industrial wastewater, and stormwater.

Water reuse The use of treated wastewater for a beneficial use, such as agricultural irrigation and industrial cooling. A process or a system for producing renewable water source, which is to be used in human activity such as agriculture, shower, flushing, and drinking, from wastewater treatment.

Definition of the Subject

Rainwater harvesting is a term covering all those techniques whereby rain is captured and used close to where it first reaches the earth [1, 2]. It includes the collecting, storing, and use of rainwater as non-potable and potable water from rooftop and other impermeable surface. The term has been also applied to arrangements to cause rainfall to infiltrate into the ground rather than run off its surface, to forms of flood control, to the construction of small reservoirs to capture runoff water so that it can be used for irrigation, and to the treatment of runoff from road and pavement surface to control nonpoint source pollutions.

Rainwater harvesting is regarded as one of the best available ways for sustainable water supply and also effective to recover natural hydrological cycles in urban area [3]. Rainwater harvesting systems provide a source of ongoing water supply and reduce reliance on other water sources [4, 5]. Moreover, rainwater harvesting not only allows an effective control of stormwater runoff by storing and infiltrating rainwater [6] but also aids the control of nonpoint source pollution in urban areas.

Introduction

Global freshwater resources are being increasingly polluted and depleted, threatening sustainable development and human and ecosystem health [7]. Water stress has been causing adverse effects and limiting economic and social development in many countries of the world. Thus, concern about the clean and safe water is a strong motivation to understand the water demands and the consequences of water use under the rapid population growth condition. It is necessary to introduce new concepts, policies, and measures for sustainable water use and management [8].

There is also a growing consensus that the way that cities and infrastructures are constructed is seriously out of step with natural hydrological and ecological processes [9, 10]. Recent environmental problems in urban areas arise from this nonresponsiveness of urban development to natural systems and processes. For example, urbanization has increased impermeable areas that have lost functions of rainwater storage and infiltration, leading to water-related disasters including urban floods and droughts [11, 12]. Increases in population and water use are resulting in increased pressures on water resources, and traditional planning concepts for water supply in urban area have a limit [13, 14]. Creating cities that are more responsive to the natural water cycle is therefore a preventative measure that will contribute to the long-term solution for water control and supply problems [10].

Rainwater harvesting is recognized as a sustainable method to overcome the problems of water shortage and urban environmental problems related to hydrological cycle. The term rainwater harvesting refers to direct collection of precipitation falling on a surface known as catchment without passing through the stage of surface runoff on land. Rainwater may be harvested from roofs, ground surfaces as well as from ephemeral watercourses and stored in physical structures or within the soil profile. Generally, rainwater harvesting involves all the activities such as collection of rainwater from roofs and other surfaces, its storage, infiltration into the ground, and subsequent use.

Almost all water used by human beings derives from rainfall. However, it is always necessary to store water, so that it can be used when rain is not actually falling, and it is often necessary to transport it from where it falls to where the user is located. Rainwater harvesting differs from other water supplies in that there is no need to transport water, since it is used where it falls as rain. Accordingly, rainwater harvesting is a means to supply water in a decentralized manner.

Rainwater harvesting provides the long-term answers to the problem of water scarcity [15]. It offers an ideal solution in areas where there is sufficient rain but inadequate ground water supply and surface water resources are either lacking or are insufficient. On the other hand, it has not been widely employed in industrialized societies that rely primarily on centralized water distribution systems. However, the role of rainwater harvesting is being reassessed in urban regions. Rainwater harvesting can be adopted in cities to provide supplemental water for the city's requirements, to increase soil moisture levels for urban greenery, to increase the ground water table through artificial recharge, to mitigate urban flooding, and to improve the quality of groundwater.

Rainwater harvesting may be implemented as a means to mitigate the adverse effect of global warming. Predictions regarding global warming could have a major effect in significantly increasing water demand in many cities [15]. At the same time, increased evaporation from reservoirs and reduced river flows in some areas may decrease the available surface water supplies. A greater uncertainty regarding yields from major reservoirs and well fields is likely to make investments in the diversification of water sources, better water management, and water conservation even more prudent in future. The role of rainwater harvesting systems as sources of supplementary, back-up, or emergency water supply will become more important, especially in view of increased climate variability and the possibility of greater frequencies of droughts and floods in many areas. This will particularly be the case in areas where increasing pressure is put on existing water resources.

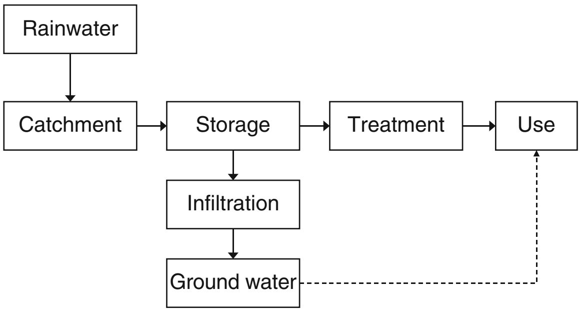
Rainwater harvesting is not a new idea. Rainwater harvesting and utilization systems have been used since ancient times and evidence of roof catchment systems date back to early Roman times [15]. Roman cities were designed to take advantage of rainwater as the principal water source for drinking and domestic purposes since at least 2000 B.C. In the Middle East, archaeological evidence of water harvesting structures appears in Jordan, Israel, Palestine, Syria, Iraq, the Negev, and the Yemen [16]. In many countries in Africa and Asia, rainwater harvesting was an important method to obtain water until the centralized water distribution system was introduced. Various technologies to harvest rainwater have been in use for over 4,000 years B.C.

There are two types of rainwater harvesting, direct rainwater harvesting and indirect rainwater harvesting. Direct rainwater harvesting implies the collection and storage of precipitated water for immediate use. Thus, it includes the rooftop water harvesting and similar systems. On the other hand, indirect rainwater harvesting means the collection and infiltration of rainfall into ground water aquifer for future use. Artificial recharge of rainwater is an example of indirect rainwater harvesting. The primary goal of direct rainwater harvesting is to supply collected rainwater as an alternative water source although excess rainwater may be infiltrated. Indirect rainwater focuses on the control of urban runoff and restoration of the hydrological cycle.

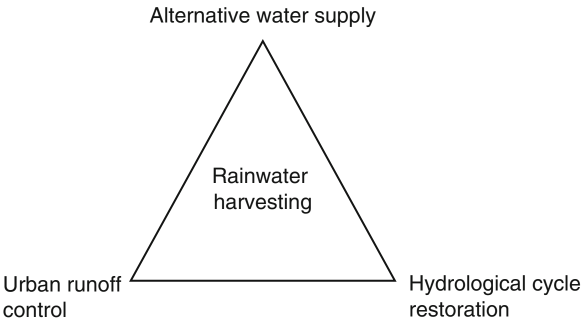
Rainwater harvesting offers a number of benefits compared with other competing technologies. It provides inexpensive supply of water and reduces stormwater runoff. Nonpoint source pollution related to stormwater can be controlled and hydrological cycles in urban areas may be recovered. In addition, rainwater needs little treatment for irrigation or non-potable indoor uses. However, there are also problems and barriers. First, the supply of rainwater is limited and unpredictable. Rainwater is often contaminated by toxic pollutants and microbes. Although rainwater is free, the initial setup costs are high. In addition, many people don't understand the importance of rainwater collection.

Rainwater harvesting system consists of several components (see Fig. 1):

- Catchment: The surface upon which the rain falls; the roof has to be appropriately sloped preferably toward the direction of storage and recharge.
- Storage: Reservoirs, tanks, etc. where collected rain water is safely stored.



Rainwater Harvesting. Figure 1
Key components of rainwater harvesting



Rainwater Harvesting. Figure 2
Basic functions of rainwater harvesting

- Infiltration: Recharging of the stored rainwater into the ground water through open wells, bore wells, or percolation pits, etc.
- Treatment: Filters and other devices to remove solids and organic material and equipment, and additives to settle, filter, and disinfect.
- Others: Conveying (the delivery system for the treated rainwater, either by gravity or pump) and other equipments.

Describing Rainwater Harvesting

As shown in Fig. 2, rainwater harvesting has three basic functions: provision of an alternative water source; control of urban runoff; and restoration of hydrological cycle that is distorted by urbanization.

Alternative Water Supply

At least one third of the population in developing countries has no access to safe drinking water, which results in major health problems due to waterborne diseases [17]. The two major water problems are inadequate supplies and insufficient treatment, which are particularly serious in developing nations [18]. Accordingly, it is essential to search for alternative approaches, including the use of decentralized water supply systems and low-cost, low-energy water technologies. Rainwater harvesting provides the long-term answers to the problem of water scarcity [15]. It helps to mitigate variation in water availability by collecting the rain and storing it more efficiently. In doing so, rainwater harvesting assures a continuous and reliable access to water. A water harvesting system collects and stores water within accessible distance of its place of use. The rainwater collected can be stored for direct use or can be recharged into the ground water.

Rainwater harvesting may offer a clean source for good quality water. This is relevant for areas where ground water or surface water is contaminated by harmful chemicals or pathogenic bacteria or pesticides and/or in areas with saline surface water. Rainwater collected using various methods has less negative environmental impacts compared to other technologies for water resources development. Rainwater is relatively clean and the quality is usually acceptable for many purposes with little or even no treatment [15]. Rainwater harvesting technologies are flexible and can be built to meet almost any requirements. Construction, operation, and maintenance are not labor intensive.

Especially, poor countries suffering or facing water shortages as a result of climate change have a massive potential in rainwater harvesting. According to a report, about a third of Africa is deemed suitable for rainwater harvesting if a threshold of 200 mm of arrival rainfall, considered to be at the lower end of the scale, is used [19]. Widely deployed, rainwater harvesting can act as a buffer against drought events for these people while also significantly supplementing supplies in cities and areas connected to the large-scale water infrastructures.

Rainwater harvesting is not only useful in poor communities but also in urban regions. Water conservation makes good economic sense and is sometimes law for private and public commercial buildings, educational facilities, and homes. There is plenty of rainfall that can be harvested and used to supplement the demands for non-potable purposes. In urban areas of the developed world, at a household level, harvested rainwater can be used for flushing toilets and washing laundry. It can also be used for showering or bathing after proper treatment. Indeed, in hard water areas, it is superior to mains water for this. Water used for non-potable purposes does not require the same level of treatment as water that must meet drinking water quality standards. In order to safely serve these needs, this water must have, however, appropriate quality depending on its specific purpose of use.

Rainwater has an advantage over reclaimed water in that it generally contains fewer pollutants and is easier to collect, and there are few regulatory barriers to outdoor uses. It can also easily be treated for more uses inside a building such as showers, hand washing, and even drinking. However, the greatest drawback to the development of rainwater harvesting systems is that the supply is not guaranteed. To compensate, storage volume must be significantly greater than that required for grey water. In addition, rising environmental pollution requires that rainwater be treated before storage and use for indoor purposes, and storage of rainwater intended for human contact can present problems with bacteria. A possible solution for this problem is to integrate rainwater harvesting and water reclamation into one system [20].

Urban Runoff Control

Rainfall inevitably creates stormwater runoff in your watershed. If precipitation occurs faster than it can infiltrate the soil or if the soil is saturated, it becomes runoff. Runoff remains on the surface and flows into streams, rivers, and eventually large bodies such as lakes or the ocean. However, urban development has dramatically changed the natural system of stormwater runoff when it rains [21]. Impervious surfaces such as driveways, sidewalks, and streets block rainfall and other precipitation from infiltrating naturally into the ground. Thus, rainfall is directed rapidly into pipes and channels instead of being absorbed by vegetation and soil. Movement of this stormwater across the soil causes urban flood and erosion.

Every time there is concentrated heavy rain, there is an overflow of water from drains, and small- and medium-sized rivers and streams repeatedly flood [2]. These conditions can often lead to an outpouring of sewage into rivers and streams from sewer outlets and sewer pumping stations, thus contaminating the quality of urban streams and rivers. Stormwater can also carry and deposit untreated pollutants, such as sediment, nutrients, and pesticides, into surface water bodies as a form of nonpoint source pollution.

Implementing a rainwater harvesting system is one way to decrease the amount of stormwater runoff and minimize the problems associated with it. By collecting and storing rainwater from roof, rainwater harvesting systems reduce stormwater runoff taking significant pressure off stormwater infrastructure and the environment. This means reduced need for investment in stormwater infrastructure (fewer, smaller pipe systems required) and reduced maintenance of stormwater infrastructure (ponds, wetlands, and pollutant traps).

Rainwater harvesting also helps to reduce the contamination of surface water. Nonpoint source pollution from urban areas represents a serious threat to water quality. Rainwater carries chemicals, nutrients, sediments, and other forms of nonpoint source into water bodies such as rivers and lakes. Rainwater harvesting allows control of stormwater runoff, thereby reducing the nonpoint source pollution.

Hydrological Cycle Restoration

Due to the rapid pace of urbanization, many of the world's large cities are facing problems related to the hydrological cycle [2]. The impermeable surfaces of concrete and asphalt structures of cities have tended to disrupt the natural hydrological cycle, and reduce the amount of rainwater permeating underground. A decrease in the area where water can penetrate speeds up the surface flow of rainwater causes water to accumulate in drains.

The hydrological cycle is closely related to heat cycle, which is also affected by impermeable surfaces. Cities can be several degrees warmer than surrounding regions due to the built environment and the concentration of human activity, a phenomenon referred to as an urban heat island [22]. Pavements have become an important contributor to this effect by altering landcover over significant portions of an urban area. The urban heat island phenomenon alters the city's natural hydrological cycle and ecological environment in a big way.

In order to achieve a comprehensive solution to this problem, new approaches to urban development are required, emphasizing sustainability and the restoration of the urban hydrological cycle. Traditionally, storm sewer facilities have been developed based on the assumption that the amount of rainwater drained away will have to be increased [2]. From the standpoint of preserving or restoring the natural water cycle, it is important to retain rainwater and to facilitate its permeation.

The natural hydrological cycle in urban areas can be rehabilitated through rainwater harvesting, which is of great importance for sustainable development of cities. Low Impact Development (LID) is a design and site development methodology that allows newly developed and/or existing sites to hydrologically mimic predevelopment conditions. For example, if a forested area is developed for commercial purposes, one LID goal would be to mimic some of the hydrological functions of trees and encourage cleansing and infiltration of site rainwater runoff. Capturing rain and encouraging it to soak into the ground as close to the location where it falls, is another goal of LID. A rainwater harvesting system can act as a large buffer, storing water for later use. The water can then be released at a slower rate via landscape watering. Stormwater retention requirements can be partially achieved by incorporating rainwater harvesting as an integral part of the design.

Technologies for Rainwater Harvesting

There are a number of types of systems to harvest rainwater ranging from very simple to the complex industrial systems. Generally, rainwater is either harvested from the ground or from a roof. The rate at which water can be collected from either system is dependent on the plan area of the system, its efficiency, and the intensity of rainfall.

Key to success in implementing rainwater harvesting is to adopt appropriate technologies. Engineering aspects including design, implementation, construction, and operation form the cornerstones of success for rainwater harvesting. Material and technique developments lead to improved and appropriate construction practices, and proper operation ensures optimal use. Water quality, sustainability, cost, and performance of rainwater harvesting systems are major areas of concern.

Catchment

The catchment area of a water harvesting system is the surface, which receives rainfall directly and contributes the water to the system. It can be a paved area like a terrace or courtyard of a building, or an unpaved area like a lawn or open ground. Temporary structures like sloping sheds can also act as catchments.

The catchment methods are listed as [15]:

- Rooftops: If buildings with impervious roofs are already in place, the catchment area is effectively available and they provide a supply at the point of consumption. The quality of collected rainwater is high compared with other catchment methods. However, the size of rooftop area is often limited and the collected rainwater may not be sufficient.
- Paved and unpaved areas: Roads and pavements and other open areas can be effectively used to harvest the runoff. The main advantage is that water can be collected from a larger area. This is particularly advantageous in areas of low rainfall. However, the collected rainwater may be contaminated by the pollutants in such surfaces.
- Storm water drains: Most of the residential colonies have proper network of stormwater drains. If maintained neatly, these offer a simple and cost-effective means for harvesting rainwater.

Runoff harvested water, which is collected from pavement surfaces and stormwater drains, should only be used for non-potable uses because of the increased risk of contamination. For in-house uses, rooftop harvested rainwater is safer for drinking purposes than the runoff harvested water.

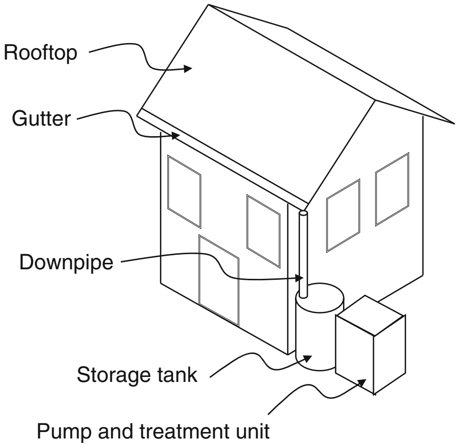
Water quality from different roof catchments is affected by the type of roof material [23]. To be suitable for rainwater harvesting, the catchment should be made of some hard material that does not absorb the rain or pollute the runoff. Thus, tiles, metal sheets, and most plastics are suitable, while grass and palm leaf roofs are generally not suitable. The surface texture affects the quantity of rainwater that can be collected from a given roof, the smoother the better [15]. In addition, roof catchment materials should resist corrosion for extended periods of time. Roof materials that could leach toxic materials into the rainwater as it touches the roof surface are recommended only for non-potable water uses (i.e., composite asphalt, asbestos, and some painted roofs).

The rainwater reaching a roof in a year can be estimated as the annual rainfall times the roof's plan area. However, some of the rainwater is typically lost to evaporation and overflow. In areas of low annual rainfall, the available roof area is not big enough to capture enough water to meet all the water needs. In this case, either the roof must be extended, or rooftop water harvesting is combined with other sources of water. In fact, getting water from more than one source is the usual practice in developing countries, and is reviving in popularity in richer countries.

One of the most challenging issues in rainwater catchment is the control of first flush effect. First flush is the initial surface runoff of a rainstorm. During this phase, water pollution is typically more concentrated compared to the remainder of the storm. Consequently, the rainwater collected in early stage of rainfall contains high concentrations of pollutants compared with that collected in later stage. The first flush effect is attributed to one or a combination of the following three processes [24]: (1) matter deposited on the catchment surface during preceding dry period is washed off by the falling rain [25]; (2) weathering and corrosion products of the catchment surface and drainage system are washed off [26]; and (3) concentrations in falling rain itself are decreasing with increasing rainfall depth due to scavenging of particles, aerosols, and gases by rain droplets [27].

To collect rainwater of high quality, rejection or treatment of first flush rainwater is essential. A first flush (foul flush) device is a valve that ensures that runoff from the first spell of rain is flushed out and does not enter the system [15]. This needs to be done since the first flush of rain carries a relatively larger amount of pollutants from the air and catchment surface. Roof washing, or the collection and disposal of the first flush of water from a roof, is of particular concern if the collected rainwater is to be used for human consumption, since the first flush picks up most of the dirt, debris, and contaminants, such as bird droppings that have collected on the roof and in the gutters during dry periods. There are a number of commercially available devices to divert or filter the first flush rainwater.

In rooftop water harvesting systems (see Figs. 3 and 4), the arrangement for leading water from the roof to the water store is usually called "guttering" or "gutters and downpipes." The gutters are open channels carrying water sideways under the edge of the roof to a point just above the water store; the downpipes are tubes leading water down from the gutters to the entrance of the water store. There are many ways of achieving the transfer of water from roof to store. However, guttering is the most popular method because it helps keep runoff water clean.

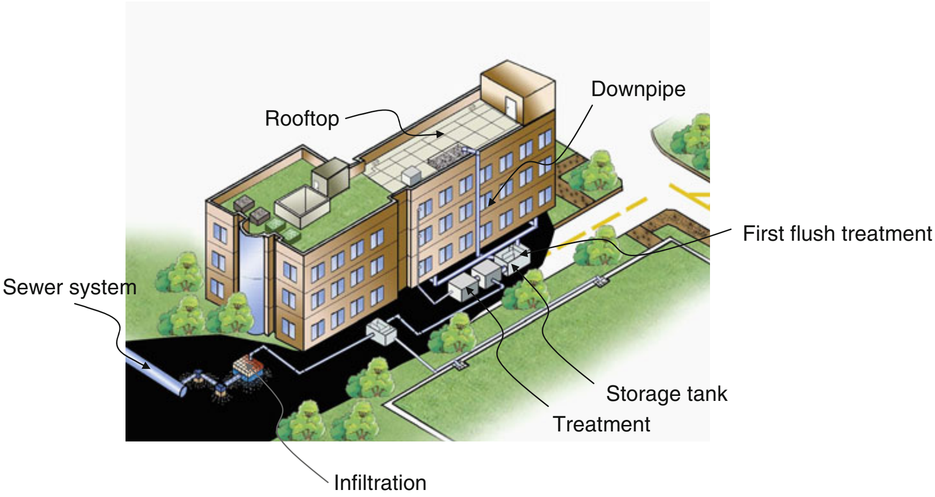


Rainwater Harvesting. Figure 3
Typical rooftop water harvesting system in a house

Storage

The storage tank or cistern generally is the most important design of a component of a rainwater harvesting system. In most cases, it is permanent and its placement should be carefully thought out. Storage tanks for collecting rainwater may be located either above or below the ground. They may be constructed as part of the building, or may be built as a separate unit located some distance away from the building. Tank location is dependent on aesthetics, climate, and soil conditions. The design considerations vary according to the type of tank and other factors. Many types of rainwater storage facilities are found in practice. Depending on the material of tanks, their characteristics are different (Table 1).

Although the level of contamination in the collected rainwater may decrease during the catchment process, it is still necessary to properly manage the rainwater storage tank to maintain high quality of water. Rainwater is usually stored for more than a week to several months before usage and thus has potential of



Rainwater Harvesting. Figure 4
Typical rooftop water harvesting system in a commercial building

Rainwater Harvesting. Table 1 Comparison of tank materials [15, 23, 28]

Tank material		Advantages	Disadvantages
Concrete		• Durable and long lasting • Ability to decrease the corrosiveness of rainwater by allowing the dissolution of calcium carbonate from the walls and floors • Install above or below ground	• Potential to crack and leak • Ongoing maintenance required • Poured in place
Plastic	Fiberglass	• Lightweight, affordable, long lasting, little maintenance • Commercially available in various sizes	• Light penetration that may promote algae growth
	Polyethylene	• Moderate price • Lightweight, low price, long lasting, little maintenance • Commercially available in various sizes • Cheap price • Easy to clean • Relatively easy to repair	• UV-degradable (Painting is required for below ground installation) • Light penetration that may promote algae growth
Metal	Galvanized steel	• Commercially available in various sizes • Moderate price	• Corrosion and rust must be lined for potable use • Only above ground use

contamination during storage. Microbial contamination is of great concern for long-term storage because it may deteriorate water quality seriously.

Infiltration

Urban rainwater cannot only be used as water supplement for daily residential use but also for improving the rate of natural rainwater infiltration and, more importantly, to rehabilitate the natural hydrological cycle. Rainwater infiltration is an approach to surface water drainage, which aims to utilize rainwater as a water supply, recharge groundwater, increase base flow levels of nearby streams, reduce flood risk, and control pollution [29]. The objective is to retain water on-site and where possible, clean it and make it available for reuse.

Infiltration is the most effective means of controlling stormwater runoff since it reduces the volume discharged to receiving waters and the negative impacts associated with it [30]. Infiltration is also an important mechanism for pollutant control. As runoff infiltrates into the ground, particulates and contaminants such as metals and nutrients are removed by filtration through the soil, and dissolved constituents are removed by adsorption. Infiltration techniques have been used for many years to control stormwater quality and flooding. Decentralized stormwater infiltration can be an additional and cost-efficient alternative resource to conventional stormwater treatment measures, which can be applied on-site close to the runoff area.

Rainwater infiltration gives rise to several benefits in relation to water quality, biodiversity, and vegetation. It is useful to reduce the amount of stormwater discharged into the sewer system and recirculate water back to the natural water cycle [29, 30]. The combination of rainwater storage and rainwater infiltration offers technical benefits and exerts positive effects on the environment and the local water balance. A hydraulic load reduction can minimize hydraulic peak loads in the sewer network significantly. In addition, infiltration of rainwater over a long period can increase and/or stabilize the groundwater reserves. The main benefits on an individual level are the complete uncoupling from the public sewer. Construction costs will be also reduced if rainwater harvesting and rainwater infiltration systems are planned together in addition to a reduction in the wastewater fees.

Infiltration is governed by two forces: gravity and capillary action. While smaller pores offer greater resistance to gravity, very small pores pull water through capillary action in addition to and even against the force of gravity. The rate of infiltration is affected by soil characteristics including ease of entry, storage capacity, and transmission rate through the soil. The soil texture and structure, vegetation types and cover, water content of the soil, soil temperature, and rainfall intensity all play a role in controlling infiltration rate and capacity. A soil hydraulic conductivity in the range of 10^{-6}–10^{-3} m/s is generally considered appropriate to achieve reasonable infiltration rates. For long-term infiltration, a fall-off due to gradual clogging of the soil must be taken into account. Moreover, the process of infiltration can continue only if there is room available for additional water at the soil surface [29, 30].

Infiltration systems are designed to capture a volume of stormwater and infiltrate this water into the ground over a period of several hours or even days. Different types of systems can be employed, which retain water safely on site, including: pervious pavements, infiltration basins, vegetated swales, ponds and wetlands, and green roofs [29].

- Infiltration basins: They are designed to capture stormwater runoff, hold this volume, and infiltrate it into the ground over a period of days. Thus, they are useful to reduce peak flows to receiving waters. Infiltration basins exhibit a high pollutant removal rate through mechanisms such as filtration, adsorption, and biodegradation. Infiltration basins may or may not be lined with plants. Vegetated systems are advantageous because they can reduce pollutants and increase the infiltration efficiency.
- Infiltration trenches: They are long, narrow gravel-filled trenches designed to infiltrate stormwater into the ground. Runoff is stored in the void space between the gravel and infiltrates through the bottom and into the soil matrix. Infiltration trenches typically capture a small amount of runoff and therefore may be designed to capture the first flush of a runoff event. Infiltration trenches efficiently remove suspended solids, particulates, bacteria, organics, and soluble nutrients.
- Vegetated swales: Swales are channels with vegetation designed to convey stormwater runoff. Vegetated swales can be applied for infiltration of not only low-polluted runoffs from roof surfaces but also high-polluted stormwater from pavement surfaces in new development areas. Except for the space requirements, they can be adapted for use in most residential, commercial, and industrial land development projects.
- Trough–trench systems: This is a combination of surface infiltration in a trough and subsurface percolation from a gravel-filled trench. The topsoil between the trough and the trench filters and purifies the infiltrating stormwater. Runoff delay is achieved through a short-term storage in the trough and long-term storage in the trench.

- Porous pavement: It is a permeable pavement surface with a stone reservoir underneath. The reservoir temporarily stores surface runoff before infiltrating it into the subsoil. Runoff is thereby infiltrated directly into the soil and receives some water quality treatment. Porous pavement often appears the same as traditional asphalt or concrete but is manufactured without "fine" materials, and instead incorporates void spaces that allow for infiltration. [29].

Treatment

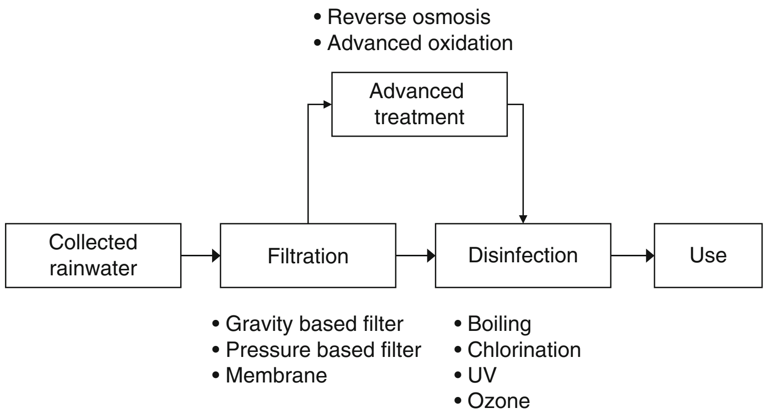
Although rainwater can provide clean, safe, and reliable water, rainwater collected in many locations still contains various pollutants. Once rain comes in contact with a roof or collection surface, it can wash many types of bacteria, molds, algae, protozoa, and other contaminants into the cistern or storage tank [31]. Indeed, some samples of harvested rainwater have shown detectable levels of these contaminants. Health concerns related to bacteria, such as salmonella and e-coli, and to physical contaminants, such as pesticides, lead and arsenic, are the primary criteria for drinking water quality analysis. Falling rain is generally free of most of these hazards. But, if the rainwater is intended for use inside the household, either for potable uses such as drinking and cooking or for non-potable uses including showering and toilet flushing, appropriate filtration and disinfection practices should be employed. Therefore, in order to ensure quality of water, the collection systems will have to be properly built and maintained and the water shall also have to be treated appropriately for intended uses.

If the rainwater is to be used outside for landscape irrigation, where human consumption of the untreated water is less likely, the presence of contaminants may not be of major concern and thus treatment requirement can be less stringent or not required at all. Depending on where the system is located, the quality of rainwater itself can vary, reflecting exposure to air pollution caused by industries such as cement kilns, gravel quarries, crop dusting, and a high concentration of automobile emissions.

Rainwater quality varies for a number of reasons [15]. While there are widely accepted standards for drinking water, the development of approved standards for water when it is used for non-potable applications would facilitate the use of rainwater sources [2]. In terms of physical–chemical parameters, collected roof water, rainwater, and urban stormwater tend to exhibit quality levels that are generally comparable to the World Health Organization (WHO) guidelines for drinking water. However, low pH rainwater can occur as a result of sulfur dioxide, nitrous oxide, and other industrial emissions. Hence, air quality standards need to be reviewed and enforced. In addition, high lead values can sometimes be attributed to the composition of certain roofing materials – thus, it is recommended that for roof water collection systems, the type of roofing material should be carefully considered. A number of collected rainwater samples have exceeded the WHO values in terms of total coliform and fecal coliform. The ratios of fecal coliform to fecal streptococci from these samples indicated that the source of pollution was the droppings of birds, rodents, etc.

Before making a decision about what type of water treatment methods to use, water should be got tested by an approved laboratory and determine whether the water can be used for potable or non-potable uses. The types of treatment discussed are filtration, disinfection, and buffering for pH control. Dirt, rust, scale, silt and other suspended particles, bird and rodent feces, airborne bacteria, and cysts will inadvertently find their way into the cistern or storage tank even when design features such as roof washers, screens, and tight-fitting lids are properly installed. Water can be unsatisfactory without being unsafe; therefore, filtration and some form of disinfection is the minimum recommended treatment if the water is to be used for human consumption (drinking, brushing teeth, or cooking). The types of treatment units most commonly used by rainwater systems are filters that remove sediment, in consort with either ultraviolet light or chemical disinfection. Advanced treatment such as reverse osmosis may be applied for special purposes (see Fig. 5).

A filter is an important part of the inflow structure of a rainwater harvesting system. Once screens and roof washers remove large debris, other filters are available, which help improve rainwater quality. Keep in mind that most filters available in the market are designed to treat municipal water or well water. Therefore, filter selection requires careful consideration. Screening,



Rainwater Harvesting. Figure 5
Schematics of rainwater treatment system

sedimentation, and pre-filtering occur between catchment and storage or within the tank [15, 23, 28].

- Gravity-based filter: This consists of construction of an underground/above ground filtration chamber consisting of layers of fine sand/coarse sand and gravel. Alternatively, only fine sand can also be used along with the gravel layer. Further deepening of the filter media shall not result in an appreciable increase in the rate of recharge and the rate of filtration is proportional to the surface area of the filter media. A unit sq.m. surface area of such a filter shall facilitate approx. 60 l/h of filtration of rainwater runoff. In order to determine the optimum size of the surface area, just divide the total design recharge potential with this figure. A system of coarse and fine screen is essential to be put up before the rainwater runoff is allowed to flow into the filtration pit. A simple charcoal can be made in a drum or an earthen pot. The filter is made of gravel, sand, and charcoal, all of which are easily available. The common types of filters are:
- Sand filters: Sand filters are commonly available, easy and inexpensive to construct. These filters can be employed for treatment of water to effectively remove turbidity (suspended particles like silt and clay), color, and microorganisms. These filters are manufactured commercially on a wide scale.
- Pressure-based filter: Pressure-based filters facilitate a higher rate of filtration in a pressurized system. Being a pressure-based system, it involves a pump of capacity 0.5–1 hp. The rate of filtration is evidently high and the quality of water is than those by gravity-based filters. They are successful for areas with larger rainwater runoff and limited space availability. Also, these filters can be put in combination with an existing tube well so as to recharge water into the same bore.
- Membrane filters: Low pressure membrane such as microfiltration and ultrafiltration can be applied for removing particles and colloids from harvested rainwater [20]. It allows direct removal of pathogens without using chemicals, although disinfection is still recommended for potable use. However, the capital and operational cost of membrane filtration is more expensive than those of other filtration systems.

To reduce the risk of microbial contamination, disinfection is needed after filtration. If the harvested rainwater is used to wash clothes, water plants, or other tasks that do not involve direct human consumption or contact, treatment beyond screening and sedimentation removal is optional. However, if the water is plumbed into the house for general indoor use such as for drinking, bathing, and cooking, disinfection is needed. While filtering is quite common in private water systems, disinfection is less common for these reasons. There are many options for disinfecting rainwater prior to use [15].

- Boiling: Boiling is a very effective method of purification and very simple to carry out. Boiling water for 10–20 min is enough to remove all biological contaminants.

- Chlorination: Chlorination is done with sodium hypochlorite or calcium hypochlorite. Chlorination can kill all types of bacteria and make water safe for drinking purposes. Chlorine is the most common disinfectant because of its dependability, water solubility, and availability. However, chlorine is disliked due to taste, fear associated with trihalomethanes (THMs), and other concerns.
- Chlorine tablets: Chlorine tablets are easily available commercially. One tablet of 0.5 g is enough to disinfect 20 l (a bucketful) of water.
- Ultraviolet (UV) light water disinfection: It kills most microbiological organisms that pass through them. Since particulates offer a hiding place for bacteria and microorganisms, pre-filtering is necessary for UV systems. To determine whether the minimum dosage is distributed throughout the disinfection chamber, UV water treatment units should be equipped with a light sensor. Either an alarm or shutoff switch is activated when the water does not receive the adequate level of UV radiation. The UV unit must be correctly calibrated and tested after installation to insure that the water is being disinfected. Several systems in US utilize ultraviolet light.
- Ozone: Ozone is a strong disinfectant that readily kills microorganisms and oxidizes organic matter in the water. Any remaining ozone reverts back to dissolved oxygen in the water. Recent developments have produced compact ozone units for home use. Since ozone is produced by equipment at the point of use with electricity as the only input, it is easier to handle than chlorine or other chemicals. Ozone can also be used to keep the water in cisterns without microbial growth.

When a disinfectant such as chlorine is added to rainwater, an activated carbon filter at the tap may be used to remove the chlorine prior to use. It should be remembered that activated carbon filters are subject to becoming sites of bacterial growth. Chemical disinfectants such as chlorine or iodine must be added to the water prior to the activated carbon filter. If ultraviolet light or ozone is used for disinfection, the system should be placed after the activated carbon filter. Many water treatment standards require some type of disinfection after filtration with activated carbon. Ultraviolet light disinfection is often the method of choice.

Currently, water quality control in roof water collection systems is limited to diverting first flushes and occasional cleaning of cisterns. Boiling, despite its limitations, is the easiest and surest way to achieve disinfection, although there is often a reluctance to accept this practice as taste is affected. Chlorine in the form of household bleach can be used for disinfection. However, the cost of UV disinfection systems is usually prohibitive.

System Design

The effectiveness of rainwater harvesting depends on various factors including climate conditions, geographic information, water usage patterns, and the major purpose of collected rainwater. Therefore, the design of rainwater harvesting systems is of primary importance.

The rate at which water can be collected is dependent on the plan area of the system, its efficiency, and the intensity of rainfall. Thus, the basic system design of rainwater harvesting systems should be determined based on the size and nature of the catchment areas, water quality of collected rainwater, and whether the systems are in urban or rural settings. Some of the systems are described below [15]:

- Simple roof water collection systems: This system is suitable for household use of rainwater. While the collection of rainwater by a single household may not be significant, the impact of thousands or even millions of household rainwater storage tanks can potentially be enormous. The main components in a simple roof water collection system are the cistern itself, the piping that leads to the cistern and the appurtenances within the cistern.
- Larger systems for educational institutions, stadiums, airports, and other facilities: When the systems are larger, the overall system can become a bit more complicated, for example, rainwater collection from the roofs and grounds of institutions, storage in underground reservoirs, treatment, and then use for non-potable applications.

- Roof water collection systems for high-rise buildings in urbanized areas: In high-rise buildings, roofs can be designed for catchment purposes and the collected roof water can be kept in separate cisterns on the roofs for non-potable uses.
- Land surface catchments: Rainwater harvesting using ground or land surface catchment areas can be a simple way of collecting rainwater. Compared to rooftop catchment techniques, ground catchment techniques provide more opportunity for collecting water from a larger surface area. By retaining the flows (including flood flows) of small creeks and streams in small storage reservoirs (on surface or underground) created by low-cost (e.g., earthen) dams, this technology can meet water demands during dry periods. There is a possibility of high rates of water loss due to infiltration into the ground, and because of the often marginal quality of the water collected, this technique is mainly suitable for storing water for agricultural purposes.
- Collection of storm water in urbanized catchments: The surface runoff collected in stormwater ponds/reservoirs from urban areas is subject to a wide variety of contaminants.

A basic goal for sizing any rainwater harvesting system is to balance the volume of water that can be captured and stored (supply), compared to the volume of water used (demand). In order to balance the system, the supply must equal or exceed the demand. This is easiest to understand if broken down on a monthly basis.

If rainfall is fairly evenly distributed throughout the year, the storage capacity may be smaller than in other areas of the country where rainfall occurs more seasonally. Storage capacity needs to be sufficient to store water collected during heavy rain events to last through dry periods. Some residences might be constrained by the size of the collection surfaces and/or the volume of storage capacity that can be installed due to space or costs.

There are many ways to determine the amount of rainfall, the estimated demand, and how much storage capacity is needed to provide enough rainwater to meet the demand. The simplest way to size the rainwater storage tank is to multiply catchment area with a certain factor (0.01∼ 0.1m). Although this method is widely used, it does not consider the usage rate, leading to an inefficient design. Therefore, a design method based on simulation is preferred. A few softwares for simulating rainwater harvesting, such as MUSIC (http://www.toolkit.net.au/Tools/MUSIC), STORM (http://www.azstorm.org), and RainCity (http://www.rainwater.re.kr), are commercially available. These softwares may consider many factors including target function, catchment type, catchment area, and economic feasibility.

Operation and Maintenance

Operation and maintenance are also important steps to ensure high efficiency of rainwater harvesting. Appropriately designed rainwater harvesting systems require little maintenance. However, like any household component, it should be checked periodically to ensure it is operating efficiently and appropriately [28]. As a matter of fact, many problems related to rainwater harvesting arise from improper operation and maintenance. There are a few general guidelines:

- Catchment: The catchment surface should be as clean as possible for use of collected rainwater. Periodic cleaning of the catchment surface is recommended.
- Storage: The storage tank should be regularly cleaned to prevent contamination by sediments and microorganisms. In rooftop water system, the cleaning frequency is once a year. However, the cleaning frequency should be determined based on the characteristics of collected rainwater. If a first flush filter is not used, tanks will require periodic cleaning to remove organic debris buildup. If a first flush filter is used, tanks will not require frequent cleaning.
- Infiltration: The infiltration system may be blocked by pollutants in rainwater. Cleaning is often inefficient to recover the initial infiltration rate. Thus, the water flowing into the infiltration system should be properly treated to reduce the amount of particles.
- Treatment: The quality of treated water should be regularly monitored.

Future Directions

Rainwater harvesting has recently gained popularity as a way to secure sustainable water resource. This is becoming widespread all over the world. Rainwater harvesting may be implemented to supply clean water in developing countries such as Africa and Asia. It may be applied in cities to lessen the load of municipal water system, control urban flood, and recover natural hydrological cycle. Rainwater harvesting has been also identified as a technology with the potential of contributing immensely as a coping mechanism for climate change and variability.

However, rainwater harvesting has some limitations. The supply from rainwater harvesting is not guaranteed. In urban areas, separate systems of rainwater harvesting are not effective to reduce runoff and recover water cycle. Moreover, the understanding on rainwater harvesting is hampered by the lack of scientifically sound principles and research.

Accordingly, technologies for rainwater harvesting will be developed in many ways in the future. Rainwater harvesting will continue to be combined with wastewater reclamation for stable supply of high-quality water. Rainwater harvesting will be systematically considered in urban design to maximize its effect on runoff control and water cycle restoration. It will be an essential part of a decentralized infrastructure for water management.

Bibliography

1. Thomas TH, Martinson DB (2007) Roofwater harvesting: a handbook for practitioners. IRC, Delft
2. UNEP (2002) Rainwater harvesting and utilisation: an environmentally sound approach for sustainable urban water management: an introductory guide for decision-makers. UNEP, Osaka
3. Kim RH, Lee S, Kim YM (2003) Development of rainwater utilization system in Korea. In: 11th IRCSA, Mexico City, Mexico
4. Bambrah GK (1993) Rainwater catchment systems and technologies – An overview. In: 6th international conference on rainwater catchmant systems “Participation in rainwater collection for low-income communities and sustainable development”, Nairobi, Kenya
5. Herrmann T, Hasse K (1998) Ways to get water: rainwater utilization or long-distance water supply? A holistic assessment. Water Sci Technol 36(8–9):313–318
6. Fewkes A (1999) The use of rainwater for WC flushing: the field testing of a collection system. Build Environ 34(6): 765–772
7. Furumai H (2008) Rainwater and reclaimed wastewater for sustainable urban water use. Phys Chem Earth, Parts A/B/C 33(5):340–346
8. Wagner W et al (2002) Sustainable watershed managment: an international multi-watershed case study. Ambio 31(1):2–13
9. Coombes P, Donovan I, Cameron C (1999) Water sensitive urban development: Implementation issues for the lower hunter and central coast. Lake Macquarie City Council, Lake Macquarie City, p 99
10. Weigert FB, Steinberg CEW (2002) Sustainable development: assessment of water resource management measures. Water Sci Technol 46(6–7):55–62
11. Kim RH (2002) Rainwater utilization for urban establishment of new paradigm. In: A joint conference with Korea society of water and wastewater and Korea society on water quality, Goyang
12. Kim RH (2000) Rainwater utilization and functional changes in building roof. Construct Technol Rev 220:13–19
13. Konig KW (2001) Rainharvesting in building. Wilo Brain, Dortmund
14. Masaru T (2002) Strategies toward buidling “Green” society. Central law publishing Co, London
15. Ray K (2005) Rainwater harvesting and utilisation. UN-HABITAT
16. Prinz D (2002) The role of water harvesting in alleviating water scarcity in arid areas. In: International conference on water resources management in arid regions. Kuwait Institute for Scientific Research, Kuwait
17. WHO (2004) Facts and figures: water, sanitation and hygiene links to health. WHO, New York
18. Murinda S, Kraemer S (2008) Exploring the potential of solar water disinfection as a household water treatment method in peri-urban Zimbabwe. Phys Chem Earth, Parts A/B/C 33:8–13
19. Calabria K (2006) Devastating rains may hold solution to Africa’s water woes: report in Agence France Presse
20. Kim R-H et al (2007) Reuse of greywater and rainwater using fiber filter media and metal membrane. Desalination 202(1–3): 326–332
21. Lampe L, Andrews H, Kinsinger K (1996) 10 Issues in urban stormwater pollution control. American City and County, pp 36–53
22. Cambridge Systematics I (1005) Cool pavement report: EPA cool pavements study – Task 5, EPA
23. Krishna HJ (2005) The texas manual on rainwater harvesting. Texas Water Development Board, Austin
24. Zobrist J et al (2000) Quality of roof runoff for groundwater infiltration. Water Res 34(5):1455–1462
25. Yaziz MI et al (1989) Variations in rainwater quality from roof catchments. Water Res 23(6):761–765
26. Zhang X et al (2002) Determination of instantaneous corrosion rates and runoff rates of copper from naturally patinated copper during continuous rain events. Corros Sci 44(9): 2131–2151
27. Zinder B, Schuman T, Waldvogel A (1988) Aerosol and hydrometer concentrations and their chemical composition during winter precipitation along a mountain slope – II. Enhancement below cloud scavenging stably stratified atmosphere. Atmos Environ 22(12):2741–2750

28. LaBranche A et al (2007) Virginial rainwater harvesting manual. The Cabell Brand Center, Salem
29. Khoury-Nolde N (2008) Rainwater infiltration. European Rainwater Catchment Systems Association
30. CSIRO (2006) Urban stormwater: best practice environmental management guidelines. CISRO, Collingwood
31. Pressman A (2007) Architectural graphic standards. American Institute of Architects. Wiley, New Jersy

Rating Systems for Sustainability

Raymond J. Cole
School of Architecture & Landscape Architecture,
University of BC, Vancouver, BC, Canada

Article Outline

Glossary
Definition of the Subject
Introduction
Environmental Assessment Methods
Sustainability Assessment
Future Directions
Acknowledgments
Bibliography

Glossary

Building Environmental Assessment Method (system or scheme) Technique that has environmental assessment as one of its core functions but may be accompanied by third party verification before issuing an overall performance rating or label.

Assessment process Use of assessment methods, including deployment by the design team and engagement of other stakeholders as the basis for making informed decisions.

Certification Third party verification and scrutiny of a performance assessment that adds to the overall credibility of the assessment process but invariably brings additional layers of constraints, bureaucracy, and costs.

Environmental (or green) assessment Assessment of resource use, ecological loadings, and indoor environmental quality.

Framework Organization or classification of environmental performance criteria in a structured manner with assigned points or weightings.

Rating (labeling) Extended output from the assessment process, typically in the form of a singular, easily recognizable designation, for example, "Gold" or "Excellent."

Sustainability assessment Assessment that expands the range of performance criteria to include social and economic considerations.

Weighting Assigning the relative significance of the environmental criteria to permit their aggregation into an overall single score.

Definition of the Subject

Voluntary building environmental assessment methods have emerged as a legitimate means to evaluate the performance of buildings across a broad range of environmental considerations – most typically resource use, ecological loadings, and indoor environmental quality. An underlying premise of these voluntary assessments is that if the market is provided with improved information and mechanisms, a discerning client group can and will provide leadership in environmental responsibility, and that others will follow suit to remain competitive.

The increase in development and application of building environmental assessment methods over the past 20 years has provided considerable theoretical and practical experience on their contribution in furthering environmentally responsible building practices. An important indirect benefit is that the broad range of issues incorporated in environmental assessments require greater communication and interaction between members of the design team and various sectors with the building industry, that is, environmental assessment methods encourage greater dialogue and teamwork. Although the developers of assessment tools continue to refine their scope, structure, and metrics, key issues are the interpretation placed on the final environmental profile or label by the "market," its significance alongside other design requirements, and the changing relationship between voluntary and regulatory mechanisms to improve building environmental performance. Moreover, while the term "sustainable" building is increasingly used and social and economic criteria are being added to the assessments, few current assessment methods have been designed and structured from the outset to embrace these considerations.

Introduction

Until the release of the *Building Research Establishment Environmental Assessment Method* (BREEAM) in 1990 [1] little, if any, attempt had been made to establish an objective and comprehensive means of simultaneously assessing a broad range of environmental considerations. Building environmental rating methods offer a means to demonstrate that a building has been successful at meeting an expected level of performance in a number of declared criteria. They provide building owners with a credible and objective means to communicate to prospective tenants the environmental qualities of the building they are leasing and, by emphasizing more demanding performance goals and the benefits over typical practice, they offer the potential for reframing expectations.

Building environmental assessment methods typically consist of three major components:

- A declared set of environmental performance criteria organized in a logical fashion – the *structure*
- The assignment of a number of possible points or credits for each performance issue that can be earned by meeting a given level of performance – the *scoring*
- A means of showing the overall score of the environmental performance of a building or facility – the *output*

Deriving a final aggregate "score" invariably requires some form of "weighting" – either implicitly or explicitly – being applied to the constituent criteria to reflect their relative significance within the overall measure of performance. This weighting process, which can profoundly influence the final overall building performance designation, has consistently represented a key part of discussions on building environmental performance. Key issues here relate to need to establish a more rational basis for the derivation of weightings, how these reflect regional priorities, and whether they should be implicit or explicit in the scoring process.

The development of assessment methods has, for the main part, been driven by the scoping and structuring of performance criteria. Although it is generally accepted that environmental criteria must be organized in ways that facilitate meaningful dialogue and application, the structuring of criteria within the assessment method is most important during the *output* of the performance evaluation, when the "story" of the performance must be told in a coherent and informative way to a variety of different recipients. Gann et al. [2] indicate that the "methods by which results are depicted has a direct bearing on how the indicators are used and understood – and by whom." The Japanese *Comprehensive Assessment System for Building Environmental Efficiency* (CASBEE) [3], for example, explicitly distinguishes between the way that performance information is organized during the assessment process and how it is transformed to communicate a variety of different outputs. It uses a variety of different output formats, providing the opportunity to tell different "stories" about a building's performance – an overall performance as well as more detailed descriptions. Moreover, CASBEE, while employing an additive/weighting approach, breaks away from the simple addition of points achieved in *all* performance areas to derive an *overall* building score, which has been the dominant feature of all previous methods. It distinguishes between the *Environmental Loading* (resource use and ecological impacts) and *Environmental Quality and Performance* (indoor environmental quality and amenities), scoring them separately to determine the Building Environmental Efficiency, that is, the ratio of *Environmental Quality and Performance* to *Environmental Loading*. As such, the structure of the CASBEE itself conveys an eco-efficiency view of assessment.

Given that assessment methods function as voluntary, market place mechanisms, ensuring that the methods are simple, practical, and inexpensive in both use and maintenance is deemed paramount. This simple characterization of building environmental issues has both positive and negative impacts on building design. For owners and design teams beginning to address environmental issues, the simplicity provides a straightforward means of discovering what is important and what is not important. As such assessment methods can be an instrument for changing design practice by identifying a new standard of performance that encourages architects and engineers to break old habits and design norms. However, building environmental methods may, indirectly, detract from a more fundamental professional commitment to environmental responsibility. Currently, design teams and owners use assessment methods to collectively establish an overall desired "score" – a BREEAM "Excellent" or "Very Good," or a US Leadership in Energy and Environmental Design (LEED®) [4] "Gold" or "Silver" rating – and then review the range of individual performance criteria to identify those seen as being attainable. There is concern that "checklist" type environmental assessment methods may drive the product and process where simply achieving a high score, with many important gains, may prove more important than aspiring to the best overall product.

Environmental Assessment Methods

Assessment implies measuring how well or poorly a building is performing, or is likely to perform, against a declared set of criteria. Most current building assessment methods attempt to measure improvements in the environmental performance of buildings relative to current typical practice or requirements and have the following general characteristics:

- Are technically framed and emphasize the assessment of resource use, ecological loadings, and health and comfort in individual buildings
- Are primarily concerned with mitigation – reducing stresses on natural systems by improving the environmental performance of buildings
- Assess performance relative to explicitly declared or implicit benchmarks and, as such, measure the extent of improvement rather than proximity to a defined, desired goal
- Assess design intentions and potential as determined through prediction rather than actual real-world performance
- Structure performance scoring as a simple additive process and use explicitly declared or implicit weightings to denote priority
- Offer a performance summary, certificate, or label that can be part of leasing documents and promotional documents

Building environmental assessment is now a distinct and important realm of research and inquiry that seeks to develop greater refinement and rigor in performance indicators, weighting protocols, and, where appropriate, the potential incorporation of Life Cycle Assessment (LCA) approaches to refine the constituent measures. Moreover, it has provided numerous side-by-side comparisons of the more notable methods and tools (e.g., BREEAM, LEED®, CASBEE, and the Australian Green Star) to illustrate areas of convergence and distinction, typically as a starting point for generating applicable methods in other regions or countries seeking to develop new assessment schemes. Within this debate, the scope of comparison and analysis is typically based only on technical content (e.g., the framework) and makes little or no reference to the organizational or market context within which the methods operate; that is, comparisons are made indiscriminately between tools and methods. This represents a serious problem since the context within which an assessment method has been designed to operate profoundly affects the effective scope, emphasis, and rigor of an assessment.

Although initially introduced to perform a specific assessment role as a means to counter-act unverified claims of building performance – "green-wash" – building environmental assessment methods play a qualitatively different role in today's context. There are several emerging issues that increasingly frame the use of building assessment methods:

- Assessment methods have moved beyond voluntary market place mechanisms. Performance thresholds in the assessment methods (e.g., LEED® Gold) are increasingly being specified by public agencies and other organizations as performance requirements, and are being considered as potential incentives for development approval, bonus density, and other concessions.
- Building environmental assessment is increasingly being recognized by banking, financial, and insurance companies as a basis for risk and mortgage appraisals and real estate valuations.

- With more widespread adoption of assessment tools, compliance with performance requirements increasingly affects associated manufacturing industries. While some industries use this as an opportunity to reevaluate production processes, some become increasingly resistive.
- The range of building types seeking certification is increasing and this, in turn, is creating the need either to develop generic systems that can recognize distinctions on an as-needed basis for specific situations or to create a suite of related methods and tools, each of which uniquely addresses a particular building type.
- The need to permit easy access to tools and methods, and to enable assessments to be made quickly and cheaply, is spurring the increased deployment of Web-based methods and electronic submission processes to attain certification.
- The aggregate effect of individual buildings has enormous consequence for community infrastructure design and operation. This, together with the inherent limitation of analyzing individual buildings as the basis to understand ecological impacts, has generated interest in creating and linking assessment methods and tools across a variety of scales.
- Increased awareness of the inevitability of climate change has extended the approach from solely one of mitigation to now embrace adaptation to changing conditions and the conscious restoration of previously degraded natural systems.

Proliferation of Building Environmental Assessment Methods

The field of building environmental assessment has matured remarkably quickly since the introduction of BREEAM, and the interim period has witnessed a rapid increase in the number of methods either in use or being developed worldwide. Within this relatively short time period, successive generations of systems have evolved as a result of accumulated experience, new conceptual insights, and theoretical propositions. Different systems have greater strengths and weaknesses than others, and later systems draw on these to include features and elements that permit more effective use. The past decade has witnessed a proliferation of building environmental assessment methods by countries worldwide for application within their respective domestic markets (see Table 1). Many of these systems have made considerable conceptual advances over early, more established methods. To date, however, with the exception of three or four systems, the number of assessed buildings in many countries using the domestic systems remains modest. This lack of traction is primarily due to the lack of the organizational and financial resources required to support the necessary educational, management, and certification programs.

There is little doubt that building environmental assessment methods have contributed enormously to furthering the promotion of higher environmental expectations and are directly and indirectly influencing the performance of buildings. Assessment methods have enjoyed considerable success and their widespread awareness has created the critical mass of interest necessary to cement their role in creating positive change. “Success” is used here in reference to the way that assessment methods have entered the parlance of the building industry rather than the number of actual “assessed” and “certified” projects – which is still relatively low compared to the total number of buildings constructed annually [5].

A number of factors have collectively generated the early momentum of assessment methods:

- The prior absence of any means to both discuss and evaluate building performance in a comprehensive way left open a distinct niche within an emerging European and North American “culture of performance assessment.”
- The simple, seemingly straightforward declaration of the requirements of a discrete number of performance measures presented a complex set of issues in a manageable form.
- By offering a recognizable structure for environmental issues, they provided a focus for the debate on building environmental performance.
- Public sector building agencies have used them as a means of demonstrating commitment to emerging environmental policies and directives.
- Manufacturers of “green” building materials and products have been given the opportunity to make direct and indirect associations with the relevant performance criteria, to support the sales of

Rating Systems for Sustainability. Table 1 Building environmental assessment methods and tools in use worldwide

Region	Country	Name	Owner/Management
Europe	France	HQE Method	HQE (*Haute Qualité Environnementale)*
	Finland	PromisE	VTT (Technical Research Centre of Finland)
	Germany	Sustainable Building Certificate	German Sustainable Building Council
	Italy	Protocollo ITACA	iiSBE Italia
	Norway	Envir. Programming of Urban Development	SINTEF (Skandinavias største uavhengige forskningsorganisasjon)
	Portugal	LiderA (Leadership for the Environment in Sustainable Building)	Instituto Superior Técnico, Lisbon
	Spain	VERDE	Spanish GBC
	Sweden	EcoEffect	Royal Institute of Technology
	Netherlands	BREEAM-NL	Dutch GBC
	UK	BREEAM (Building Research Establishment Environmental Assessment Method)	Building Research Establishment
	Europe	LEnSE (Label for Environmental, Social & Economic building)	Belgian Building Research Institute
N. America	United States	LEED® (Leadership in Energy and Environmental Design)	United States GBC
		GreenGlobes	Green Building Initiative
	Canada	LEED-Canada	Canada GBC
		GreenGlobes	ECD Canada
	Mexico	SICES	Mexico GBC
Asia	China	GHEM (Green Housing Evaluation Manual)	Ministry of Construction
		GOBAS (Green Olympic Building Assessment Scheme)	Ministry of Science & Technology
		ESGB (Evaluation Standards of Green Building)	Ministry of Housing and Urban-Rural Construction
	Hong Kong	BEAMPlus	HK-BEAM Society
		CEPAS (Comprehensive Environmental Performance Assessment Scheme)	HK Building Department
	India	TERI-GRIHA (Green Rating for Integrated Habitat Assessment)	TERI (The Energy & Research Institute)
		LEED-India	Indian GBC
	Japan	CASBEE (Comprehensive Assessment System for Building Environmental Efficiency)	Japan Sustainable Building Consort.
	Korea	GBCC (Green Building Certification Criteria)	Korean Korea Institute of Energy Research

Rating Systems for Sustainability. Table 1 (Continued)

Region	Country	Name	Owner/Management
	Singapore	Green Mark	Singapore Building & Construction Authority (BCA)
	Taiwan	EEWH (Ecology, Energy, Waste and Healthy)	ABRI (Architecture and Building Research Institute)
	Vietnam	LOTUS	Vietnam GBC
Southern H.	Australia	Green Star	Australian GBC
		NABERS (National Australian Building Environmental Rating Scheme)	
	Brazil	LEED-BRAZIL	GBC Brazil
		HQE	Fundação Vanzolini
	New Zealand	Green Star NZ	New Zealand GBC
	S. Africa	Green Star SA	South African GBC
		SBAT (Sustainable Building Assessment Tool)	CSIR (Council for Scientific and Industrial Research)
Generic		GBTool/SBTool	iiSBE (International Initiative for a Sustainable Built Environment)
		SPeAR (Sustainable Project Assessment Routine)	Ove Arup Ltd.

GBC – Green Building Council

"high" performance products. The range of environmental goals has led to the engagement of diverse professional expertise early in the design process with collaborative decision-making processes emerging.

Because of this momentum, building environmental assessment methods have dwarfed all other mechanisms for instilling environmental awareness within the building industry. Indeed, over the past decade they have been positioned as the most potent mechanism for affecting change.

Sustainability Assessment

The majority of current building environmental methods assess *environmental* performance improvements *relative* to typical practice, either implicitly or explicitly. They are also set within the parlance of marketing practice, and the intentions for such systems and the individual assessment credits that comprise them are understandable by all the stakeholders. Since its introduction in 1987, "sustainability" has emerged as a widely held and necessary overarching notion to frame and guide all future human activity and enterprise. An emerging debate on building assessment relates to shift from environmental or "green" performance to this larger goal of sustainability.

The need to develop methods of deliberation and decision making that actively engage the relevant interests of stakeholders will become increasingly important to infuse sustainability considerations into day-to-day conduct and practice. Robinson [6] suggests that the "power of sustainability lies precisely in the degree to which it brings to the surface these contradictions and provides a kind of discursive playing field in which they can be debated" and subsequently encourage the "development of new modes of public consultation and involvement intending multiple views to be expressed and debated." Again, the parallel debate in building environmental assessment is becoming increasingly evident. Kaatz, Root, and Bowen [7], for example, advocate the implementation of a broader participatory approach in building assessment, including consideration of process design, definition of desired outputs and outcomes, etc.

Sustainability relates to a suite of concepts and embraces notions other than environmental performance, and as such is open to wider interpretation. However, it has two implicit requirements. Firstly, within the constituent dimensions of sustainability – environmental, social, and economic – is a responsibility to inter- and intra-generational equity. Secondly, sustainability, and any discussion of it, requires thinking long-term and assuming responsibility to the future. However, the ideal of the principles of sustainability is one thing, their assimilation and practice within the building industry is somewhat different. Whereas it is possible to define "green" and even "greener" as well as the incremental process for improving performance, it is currently difficult to envision and articulate a sustainable future – either in general terms or as related to the configuration of human settlement. It is therefore difficult to have confidence in the design of effective assessment systems when it is not possible to link them to final results. Moreover, short-sighted economic priorities and gains have always compromised environmental and social considerations in building.

Although the building development industry is fundamentally risk averse, this tendency is not evident in either recognizing or responding to larger environmental issues. Despite powerful arguments on the importance of environmental issues and evidence on the multiple benefits of early adoption of higher performance standards, the construction industry procrastinates in making changes that are perceived to increase initial cost. While the fundamental idea of economic sustainability relates to long-term and shared economic benefit, the short-term mind-set of the building industry is using the language of sustainability to maintain the status quo. What is increasingly evident at the level of the individual building is the notion of "economic sustainability" being appropriated as simply dealing with the costs of building production, and thereby relegating environmental performance issues to being included only if they make economic sense.

Sustainability Assessment Methods

The Living Building Challenge (LBC) [8], launched in August 2006, is emerging in North America as a recognized and complementary performance aspiration to LEED®. Rather than permitting choice of credits to attain an overall performance score, the LBC requires 20 demanding performance requirements – including net-zero energy – to be met before the designation of Living Building is granted. However, while it references natural systems and uses a flower/petals metaphor, there is no recognizable organization of the issues based on ecological or systems theory. Similar to LEED® and the majority of other current assessment methods, the structure is simply a list of required performance requirements set within a defined set of categories.

A number of assessment tools have been introduced that expand on the range of performance issues to explicitly include social and economic criteria and thereby attempt to provide a measure of "sustainable" performance:

- *Ove Arup's Sustainable Project Assessment Routine SPeAR®* [9]: Functions as a project assessment methodology within Ove Arup's consulting projects to enable a rapid review of the sustainability of projects, plans, products, and organizations. Performance criteria are organized in four general sustainability categories: *Environment (*Air Quality, Land Use, Water, Ecology and Cultural Heritage, Design and Operation, and Transport); *Natural Resources* (Materials, Water, Energy, Land Utilization, and Waste Hierarchy); *Economic* (Social Benefits and Costs, Transport, Employment/Skills, Competition Effects, and Viability); and *Societal* (Health and Welfare, User Comfort/Satisfaction, Form and Space, Access, Amenity, and Inclusion). The SPeAR® diagram combines, in a graphical format, the diverse issues that need to be considered in sustainable design, including social, economic, natural resource, and environmental issues, acknowledging both negative and positive results. SPeAR® can be used to highlight areas where a project/design/development performs poorly in terms of sustainability principles, and to identify opportunities to optimize performance, integrate best practice, or utilize new technology. SPeAR® provides a basis for evaluating a project's sustainability performance not in comparison with that of other buildings, but relative to strengths and weaknesses within a particular context. This permits a greater level of subjectivity in the definition of performance criteria and their interpretation during scoring.

- *iiSBE's Sustainable Building Tool (SBTool)* [10]: The current version of *SBTool* is arranged in seven Performance Issues: Site Selection, Planning, and Development; Energy and Resource Consumption; Environmental Loadings; Indoor Environmental Quality; Service Quality; Social and Economics Aspects; and Cultural and Perceptual Aspects. Whereas previous versions had focused on three distinct building types – Office buildings, Multi-Unit Residentials, and Schools – the current version allows a more generic description of buildings with multiple occupancies.
- *South African Sustainable Building Assessment Tool (SBAT)* [11]: Since the social and economic concerns in developing countries are far more pressing than those in developed countries, domestic constraints on environmental progress are therefore qualitatively different. SBAT explicitly introduces performance criteria that acknowledge social and economic issues [12]. A total of 15 performance areas are identified, equally divided within the overarching sustainability framework of environmental, social, and economic categories, each described through 5 performance criteria. Further, SBAT considers how it could become an integral part of, and subsequently influence, the building production process by relating its application to a nine-stage process based on the typical life cycle of a building: Briefing, Site Analysis, Target Setting, Design, Design Development, Construction, Handover, Operation, and Reuse/Refurbish/Recycle.
- *German Sustainable Building Council's Certificate Program* [13]: Five general sustainability "quality" categories are assessed and they form the overall aggregate building score: Ecological; Economic; Socio-cultural and Functional; Technical; and Process. A sixth quality with six sub-criteria – location – is evaluated and presented separately. Criteria within these performance areas are evaluated individually and aggregated to determine an overall performance designation of gold, silver, or bronze.

Robinson [6] suggests that if sustainability is to mean anything, "it must act as an integrating concept" and will require "new concepts and tools that are integrative and synthetic, not disciplinary and analytic; and that actively creates synergy, not just summation." When judged against these criteria, most current assessment methods and tools are left wanting in their ability to provide either insights or effective guidance on sustainability. A sustainable building is likely to be judged by the way that various systems fulfill multiple functions and, indeed, it is typically only possible to achieve high environmental performance within demanding cost and time constraints through the creative integration of systems. Similarly, while the three domains of environmental, social, and economic are typically used to frame sustainability, it is their points of intersection that are equally critical, that is, the ways and extent to which they positively or negatively influence each other. Simply adding social criteria to the current mix of environmental performance measures may not necessarily expose the way that one influences and is influenced by others. It can only do so if the method or tool is used as part of the deliberations between various stakeholders, that is, synergies are achieved through active, cross-disciplinary use of the tool, rather than simply the structure of the tool itself. That stated, it is important to ensure that environmental and social goals are not, yet again, compromised within this process.

Cross-Scale Assessment At a more pragmatic level, the notion of assessing the sustainability of an "individual" building is clearly problematic. Although one can assess the relative extent to which measures have been taken in an individual building to reduce resource use and ecological loadings, it is not possible to identify their effectiveness at addressing or furthering economic and social progress by relating them to regional or national indicators. Indeed, Gibberd suggests that "there is no such thing as a sustainable building – only buildings that enable people to live and work in sustainable ways" [14]. Assessing the ways and extent that buildings contribute progress toward sustainability will require understanding linkages across a range of scales.

Existing environmental assessment methods were primarily conceived to assess *individual* buildings, and performance issues are bounded by those factors that influence and are influenced by them. Many of the major building environmental assessment methods offer a suite of products each targeted at a specific building type or situation. The sequence in the development of assessment methods is important in revealing the increasing acknowledgment of a broader context. The majority began with a version for new office buildings and then subsequently expanded the range of products to include existing office buildings, multi-unit residentials, and then other broader applications – schools, homes, etc. Several existing systems have recently introduced versions that address a broader context, for example, USGBC's *LEED for Neighbourhood Development (LEED-ND®), CASBEE for Urban Development (CASBEE-UD)* and *BREEAM Communities*. The fact that these were developed *after* gaining experience with assessing individual buildings is remarkably telling – development has been from the scale of individual buildings upward to that of a neighborhood scale rather than setting building performance within the overarching context of a neighborhood, community, or city.

Future Directions

There are several emerging trends that will shape the future design, roles, and use of assessment tools.

Voluntary and Regulatory Mechanisms

Given the pressing time-scale of anticipated significant climate change, it is difficult to imagine that a sustainable system of production and consumption will emerge from simply tweaking current practice. A key issue, therefore, lies in the considerable difference between the levels of change that the scientific community is advocating and those that are socially and politically acceptable. Similar arguments relate to the difficulties of making significant leaps forward in achieving widespread sustainable building practices within the acknowledged conservative and cost-sensitive context of the building industry. The majority of current "green" environmental assessment methods are voluntary in their application and have the primary objective of stimulating market demand for buildings with improved environmental performance. Indeed, the "acceptance" of current assessment methods currently derives largely from their voluntary application. However, the voluntary nature of existing methods significantly compromises both their comprehensiveness and rigor. Voluntary building environmental assessment methods must serve two conflicting requirements – they must function as an objective and sufficiently demanding metric to have credibility within the environmental community, while simultaneously being attractive to building owners who wish to have something positive to show for *any* effort that they have placed on environmental performance. Satisfying these twin requirements invariably compromises both the number of criteria that are assessed, where the benchmarks are set before performance points are earned and what is presented as the most demanding performance target.

Higher environmental performance requirements are increasingly being mandated bringing into question the ways that voluntary assessment methods will have to be cast within a broader array of mechanisms for creating necessary change. Here, the current success that assessment methods enjoy within both research and practice has potentially adverse consequences. By being almost the sole focus of the debate, too much expectation may be being placed on their ability to create the necessary change.

Given the practical (and incentive) constraints on setting demand targets and dependency on market acceptance, it is uncertain whether voluntary mechanisms will ultimately be sufficient to create the necessary improvements in environmental performance of buildings needed to meet broader national environmental or sustainability targets. As such, the relationship between building environmental assessment methods and other change instruments, both regulatory and incentive based, will gain in importance [15]. Historically, regulation provided minimal acceptable performance requirements, and the voluntary mechanisms offer the complementary high performance aspiration. Recently, the mandates of far reaching performance requirements such as carbon neutrality will profoundly change these roles. In Europe, for example, demanding energy and carbon emission standards for buildings are now being introduced requiring phased reductions to net-zero energy performance [16, 17]. In North America, the recent development of ASHRAE 189.1-2009 [18] jointly by the American Society of Heating, Refrigeration, and Air Conditioning Engineers (ASHRAE), the Illuminating Engineers Society (IES), and the US Green Building Council (USGBC) sets green building performance criteria within a regulatory framework.

Achieved Performance

The assessment of building environmental performance of new buildings is typically made at the design stage and based on default patterns of occupant behavior and building operation. There is sufficient evidence to show that a building's performance in use is often markedly different from that anticipated or predicted during design and this discrepancy has initiated a shift toward basing assessments on achieved performance.

The owners and developers of any of the major assessment systems are actively seeking data on the actual performance of buildings, particularly energy and water use, and energy-related emissions. The Canada Green Building Council (CaGBC), for example, adopted the 2030 Challenge which aims for net-zero buildings by 2030 [19] and has set targets of 50% measured reductions in energy and water use from a 2005 baseline for 100,000 buildings and 1,000,000 homes by 2015. These targets cover existing and new building stock, and represent 20% of all buildings in Canada, two-thirds of total building space, and 10% of all Canadian homes. To meet these targets, the CaGBC introduced the Green Building Performance Initiative in 2008 that involves performance benchmarking, auditing, action planning and verification have informed the development and testing. GREEN UP – Canada's Building Performance Program – is providing online tools, standards, and continuously updated performance information to building owners and managers across Canada [20]. The current program is addressing performance in energy and water use, and energy-related emissions, using monthly utility billing as the data source.

Building Valuation

The need to establish a business case for the development of "green" or/and sustainable commercial properties within the real estate industry has paralleled the technical development and application of building environmental assessment methods [21]. Although the possible capital cost premiums associated with attaining higher building environmental performance has been a recurring issue over the past 20 years, the emphasis of these economic considerations has changed considerably. Initially, the business case was framed around the added benefit and reduced revenue costs to the building owner. Today, however, the business case is increasingly rooted in the added value associated with higher building environmental performance and the demonstration that green buildings may be "worth more" to investors, owners, and tenants [22].

Whereas the cost arguments have consistently referenced building environmental assessments, for example, the cost of LEED [23, 24], very little attention has been directed at connecting green rating to value. CASBEE is the first system to introduce a version explicitly linking building environmental performance assessment with real estate appraisal. CASBEE for Property Appraisal [25] is an "appraisal support tool that measures the impact degree of [design for the environment] on the property value" that when widely applied will significantly increase the demand for green buildings.

Regionalization and Standardization

The past decade or so has witnessed many countries worldwide now either having or in the process of developing domestic systems. This carries the implicit expectation for domestic systems to encourage green building practices appropriate to their specific climatic and cultural contexts. Moreover, many of these systems include innovative conceptual advances over the earlier more established methods. With the exception of iiSBE's *SBTool* and the more recent *LEnSE* project [26] that have generic core frameworks and criteria but were designed from the outset to permit regional customization, the majority of recognized assessment methods are country or context specific. All assessment tools carry the values and priorities of their authors, either implicitly or explicitly, raising questions regarding the ways and extent that – without significant adaptation – they can be meaningfully adopted by other countries. Moreover, the organizational context in which assessment methods reside, that is, who owns and manages them, is critical in terms of the credibility of the method for the broad range of industry and client stakeholders, and the human and financial resources available to maintain and implement the method. These attributes are equally significant in the marketing and widespread use of systems internationally.

Similar to the necessity and value of developing standardized Life Cycle Assessment protocols for building materials and products, that is, universal criteria for establishing boundary conditions, data quality, etc., there is increased activity in defining standardized requirements for building assessment methods (e.g., ISO TG59/SC17: Sustainability in Building Construction) [27].

Given the proliferation of domestic assessment systems worldwide over the past 15 years, a number of developments are now pushing toward increased standardization. A primary driver for this development is organizations seeking a common international vocabulary for building environmental assessment that can then facilitate communication between stakeholders and inter-building and inter-country comparisons [28]. A recent development in Europe has been the establishment of the *Sustainable Building Alliance* (SBA) [28] and the *International Sustainability Alliance* (ISA) [29]. The French *Centre Scientifique et Technique du Bâtiment* (CSTB), UK *Building Research Establishment* (BRE), and others launched the *Sustainable Building Alliance* (SBA) in April 2008 with the aim of establishing common metrics for key issues so as to provide transparency between rating systems while, importantly, still recognizing regional and national differences.

The BRE initiated the International Sustainability Alliance (ISA) in late 2009 to "drive the development of common international standards for real estate" by providing an international governance structure that "join forces with international companies, Green Building Councils, research institutes and other stakeholders in the real estate chain toward an international sustainability standard for the built environment" [30]. ISA stated goals include:

- Driving toward one single European certification standard, adaptable to local market conditions
- Expanding on the current BREEAM system and creating a third Generation System for the market and industry

Methods are owned and operated by a wide range of private and public sector organizations. However, one of the most significant developments regarding the organizational context for building environmental assessment methods is the increase in the number of *Green Building Councils* [31], their linkage through the *World Green Building Council (WorldGBC)* [32], and their sharing experiences with using assessment methods. Two aspirations with the World GBC's mission are to ensure Green Building Councils are successful and have the tools necessary to advance and to support effective green building rating systems. The ways and extent to which the various member Councils favor certain assessment tools remain uncertain at this time.

Branding

An often-stated role and expectation is that the widespread adoption of assessment methods could ultimately transform the market in its expectation and demand for buildings with higher environmental performance. Almost all current building environmental assessment methods are voluntary in their application and have the primary objective of stimulating market demand for buildings with improved environmental performance. This has been accompanied by the notion of "branding" the names of LEED® and BREEAM to building owners, purchasers, and lessors to make them synonymous with high levels of environmental performance and companion mechanisms such as the LEED Accredited Professional program to "transform" expectations and valuation of skills and knowledge of design professionals. While "branding" the assessment methods domestically is obviously a necessary process for their promotion and adoption, there is also a clear shift to the support for a few international "brand-name" systems. The primary drivers here include:

- Multinational companies who have building/development projects worldwide and who are expected to adhere to numerous national environmental assessment methods
- Corporations/companies that need to acquire green buildings when they are operating internationally to fulfill their corporate sustainability requirements
- Manufacturers of green building products who are expected to adhere to numerous national environmental assessment methods

BREEAM and LEED® are two of the oldest and most widely internationally recognized systems and, by virtue of this, have the greatest presence outside their countries of origin. They are also being increasingly positioned as being in competition within a global market [33, 34].

BREEAM Abroad Early in its development, there was a declared aspiration that BREEAM would have an international presence. In 1997, Doggart and Baldwin reported that "BREEAM type schemes have now been developed in other countries and regions, such as Hong Kong and Canada" and that "BREEAM versions are also being developed in Denmark, Norway, Australia, New Zealand and USA" [35]. While this did not materialize, the need for some level of regional adaptation was also explicitly referenced.

BRE Global now presents BREEAM as "the world's leading environmental assessment method for buildings and now communities" [36] with over 110,000 buildings certified and over half a million registered for certification [37]. BREEAM is currently having influence internationally in terms of using *BREEAM Bespoke International* or country-specific versions. Versions of BREEAM have been created for Europe and the Gulf, a country-specific version for the Netherlands, and Memorandum of Understanding has been signed between BRE Global and a group in Spain and with the Russian and Turkish GBCs, for the introduction into those respective countries.

LEED Abroad LEED® is having influence beyond the USA in terms of various countries adapting the system for their own markets or through overseas owners having buildings assessed through the USGBC:

- The Canadian and Indian Green Building Councils have created adaptations of LEED®, while groups in Brazil, Argentina, Italy, and a dozen other countries are developing/using adaptations of LEED® [38].
- The recent creation of the Green Building Certification Institute (GBCI) – established with the support of the USGBC – allows for "balanced, objective management of the LEED Professional Accreditation program, including exam development, registration, and delivery." As of January 2010, the GBCI's website records that LEED Accredited Professions are currently in 84 countries worldwide, the largest being: USA, 126,750; Canada, 1,314; China (including Hong Kong), 936; UAE, 622; UK, 280; India, 267; South Korea, 212; Mexico, 106 [39].

While the picture is not entirely clear regarding the extent of use of the BREEAM and LEED® internationally, with registered/certified projects in 103 countries the LEED® "brand" appears to currently have the greater uptake globally. Domestic and international property developers appear to be becoming increasingly interested in certifying their buildings to LEED® in order to attract these companies to move in. It is anticipated that as more and more companies move into the China market, the demand for LEED® certified buildings will further increase.

Building environmental assessment and labeling programs are considered one of the most potent and effective means to both improve the performance of buildings and transform market expectations and demand. Indeed, the environmental performance assessment of buildings is now a major business, with significant revenues generated through the certification process, licensing of the systems, training and education, and the accrediting of professionals. Many countries around the world clearly have domestic methods that will remain the sole or dominant system within their respective markets, for example, Green Star in Australia and New Zealand, CASBEE in Japan, and Green Mark in Singapore. However, there are many other countries and regions where BREEAM, LEED®, and other systems will expand their presence over the next decade as a result of increased demand and active promotion. Market forces and "branding" will, in the fullness of time, invariably play a role in dictating the extent to which voluntary systems become de facto international approaches. While there are numerous benefits for consistency in how certain performance measures are defined and evaluated, such as energy and carbon emissions, the ways and extent that methods both recognize and accommodate cultural and contextual differences when deployed internationally will remain decisive in shaping a positive outcome.

Given the increased deployment of the major systems internationally and the significance of global building practices, it is also highly likely that there will be some level of harmonization between the major tools to enable international benchmarking, especially in relation to greenhouse gas emissions, that is, the introduction of a commonly recognized standard/metric for measuring and recording energy use and carbon emissions, which could be adopted by multiple tools.

The benefits of assessment methods have been considerable: providing more comprehensive performance goals for buildings, supporting collaborative design processes, changes in economic investment, changes in industry, international collaboration, etc. They will continue to evolve in response to a host of current challenges such as ensuring that delivered performance matches design intentions, accommodating community and regional goals and priorities, and supporting a more comprehensive integration of human and natural systems.

Acknowledgments

In addition to the specific references, the content of this article draws on following previous published works:

Cole RJ (2005) Building Environmental Assessment Methods: Redefining Intentions and Roles, Building Research & Information, 35:5, pp 455–467

Cole RJ (2006) The Coexistence of Building Environmental Assessment Methods in Common Markets, Building Research & Information, 34:4, pp 357–371

Cole RJ (2009) Chapter 18: Environmental Assessment: Shifting Scales, IN: Designing High-Density Cities for Social & Environmental Sustainability, Ed: Edward Ng, Earthscan Press, pp 273–282

Cole RJ (2010) Building Environmental Assessment in a Global Market, International Journal of Sustainable Building Technology and Urban Development, 1:1, pp 11–14

Bibliography

1. Baldwin R, Leach SJ, Doggart J, Attenborough M (1990) BREEAM version 1/90: an environmental assessment for new office designs. Building Research Establishment, Garston
2. Gann DM, Salter AJ, Whyte JK (2003) The design quality indicator as a tool for thinking. Build Res Inform 31(5):318–333
3. CASBEE, Comprehensive assessment system for building environmental efficiency, Japan Sustainable Building Consortium Corps
4. Leadership in energy and environmental design (LEED) Green building rating System, US green building council
5. Cole RJ (2005) Building environmental assessment methods: redefining intentions and roles. Build Res Inform 35(5):455–467
6. Robinson J (2004) Squaring the circle? some thoughts on the idea of sustainable development. Ecol Econ 48: 369–384
7. Kaatz E, Root D, Bowen P (2004) Implementing a participatory approach in a sustainability building assessment tool. In: Proceedings of the sustainable building Africa 2004 conference, Stellenbosch, South Africa (CD Rom, Paper No. 001), 13–18 Sept 2004
8. LBC (2010) Living building challenge version 2.0. International Living Building Institute, Seattle, WA
9. Sustainable Project Assessment Routine (SPeAR®) http://www.arup.com/environment/feature.cfm?pageid=1685
10. iiSBE (International Initiative for a Sustainable Built Environment) (2010) Sustainable building tool. http://www.iisbe.org
11. Gibberd J (2001) The sustainable building assessment tool – assessing how buildings can support sustainability in developing countries. Continental shift 2001 – IFI international conference, Johannesburg, South Africa, 11–14 Sept 2001
12. Gibberd J (2005) Paper 04–001, Assessing sustainable buildings in developing countries – The Sustainable Building Assessment Tool (SBAT) and The Sustainable Building Lifecycle (SBL). The 2005 world sustainable building conference, Tokyo 27–29 Sept 2005
13. German Sustainable Building Council's Certificate Program. http://www.gesbc.org/
14. Gibberd J (2001) The opinion of Gibberd. Sustain Build (3):41
15. Cole RJ (1999) Building environmental assessment methods: clarifying intentions. Build Res Inform 27(4/5):230–246
16. European Parliament (2009) Press release. www.europarl.europa.eu/sides/getDoc.do?pubRef=–//EP//TEXT+IM-PRESS+20090330IPR52892+0+DOC+XML+V0//EN&language=EN. Last accessed Sept 2009
17. Department for Communities and Local Government (2006) Code for sustainable homes: a step-change in sustainable home building practice. www.communities.gov.uk.
18. ASHRAE/USGBC/IES, Standard 189.1-2009, Standard for the design of high-performance green buildings (Except low-rise residential buildings), American society of heating, Ventilation and air conditioning engineers, Atlanta, Georgia
19. 2030 Challenge, The 2030 Challenge was issued by Architecture 2030, a non-profit organization. www.architecture2030.org.
20. Jarvis IA (2009) Closing the loops – how real building performance data drives continuous improvement. Intell Build Int 1(4):269–276
21. Lorenz D, Lützkendorf T (2008a) Sustainability in property valuation – theory and practice. J Property Investment Finance 26(6):482–521
22. Sayce S, Sundberg A, Mohd A (2009) Sustainable property: a premium product? A working paper, Paper presented at ERES Conference 2009, Stockholm, 24–27 June 2009
23. Kats, Gregory (2003b) The costs and financial benefits of green building: a report to california's sustainable building task force. California: Capital E. October. www.usgbc.org/Docs/News/News477.pdf

24. Matthiessen Lisa F, Peter M (2004) Costing green: a comprehensive cost database and budgeting methodology. Davis Langdon Adamson, Los Angeles
25. CASBEE for property appraisal, Japan Sustainable Building Consortium Corps. December 2009
26. LEnSE (Methodology Development towards a Label for Environmental, Social and Economic Buildings) (2006). "Stepping Stone's 1,2 & 3," LEnSE Partners, November 2006
27. ISO CD 21931, Framework for methods for assessment of environmental performance of construction works – Part 1 – Buildings, international organization for standardization, standard under development ISO TG59/SC17: sustainability in building construction
28. Visier JC (2009) Common metrics for key issues. SB alliance annual conference Paris, 5 Nov 2009
29. Sustainable Building Alliance. http://www.sballiance.org/
30. International Sustainability Alliance. http://www.bre.co.uk/page.jsp?id=2019
31. Green Building Councils. Country based members of the World Green Building Council – either established members, emerging members or prospective members
32. World Green Building Council founded in (1999) http://www.worldgbc.org/
33. Julien A (2009) Assessing the assessor: BREEAM VS LEED. Sustain Mag 9(6):30–33
34. Online (2009) BREEAM V LEED, Onoffice Magazine, May 2009
35. Doggart J, Baldwin DR (1997) BREEAM international: regional similarities and differences of an international strategy for environmental assessment of buildings. In: Proceedings second international conference: buildings and the environment, Paris, pp 83–90
36. BRE Global (2010) http://www.bre.co.uk/page.jsp?id=1763
37. BRE Global (2009) BREEAM – The environmental system for buildings around the world
38. LEED International Program. http://www.usgbc.org/DisplayPage.aspx?CMSPageID=2346
39. Numbers of LEED Accredited Professionals worldwide.http://www.gbci.org/DisplayPage.aspx?CMSPageID=113

Rechargeable Batteries, Separators for

Shriram Santhanagopalan, Zhengming (John) Zhang
Celgard, LLC, Charlotte, NC, USA

Article Outline

Glossary
Definition of the Subject
Introduction and Scope
Separator and Batteries
Separator Classification
Separators for Lithium-Ion Batteries
Separator for Lithium Polymer Batteries
Separator for Lithium-Ion Gel Polymer Batteries
Separator for Aqueous Batteries
Mathematical Modeling of Separators
Summary
Future Directions
Acknowledgments
Bibliography

Glossary

Dry process A process used to make a separator, which involves melting a polyolefin resin and extruding it into a film, thermal annealing to increase the size and amount of lamellar crystallites, and stretching to form tightly ordered micro pores.

Gurley A measure of time taken for a predetermined quantity of air (or other specified fluid) to permeate across a porous membrane.

Nonwoven separators Sheet, web, or matt of directionally or randomly oriented fibers, bonded by friction, and/or cohesion, and/or adhesion excluding paper and products which are woven, tufted, stitch-bonded incorporating binding yarns or filaments, or felted by wet milling.

Separator A porous membrane placed between electrodes of opposite polarity, permeable to ionic flow but preventing electronic contact between the electrodes.

Shutdown separator A separator that responds to increasing heat within a cell, by closure of the pores, such that it stops ionic flow between the anode and cathode, during a thermal event.

Wet process A process used to make a separator that involves a solvent.

Definition of the Subject

The battery industry has undergone tremendous growth over the last couple of years – both in terms of technological growth as well as in the variety of applications. The need to optimize battery materials to meet the increasing demand for energy as well as to extend the operating range continues to be a challenge – more so now than ever before. Correspondingly, the demand for novelty in separator membranes to match the newer battery chemistries and geometries continues to grow.

A separator is a porous membrane placed between electrodes of opposite polarity, permeable to ionic flow but preventing electronic contact between the electrodes [1, 2]. A variety of separators have been used in batteries over the years – from cedar shingles and sausage casing to present-day microporous flat sheet membranes made from polymeric materials. Their main function, however, continues to be the same – to keep the positive and negative electrodes apart. They should be very good electronic insulators and have the capability of conducting ions by either intrinsically being an ionic conductor or by soaking an electrolyte. They should minimize any processes that adversely affect the electrochemical energy efficiency of the batteries.

The separator is among the components that have sustained maximum growth in the battery industry with minimal changes to the material ingredients. Not much attention has been given to separators even in publications reviewing batteries [1–7]. The number of reviews on separators [8–17] is minimal compared to those on cell fabrication, their performance, and application in real life. Kinoshita et al. have reviewed different types of membranes/separators used in different electrochemical systems, including batteries [9]. Zhang presented a more recent review of the various separators used in liquid electrolyte systems [10]. This chapter is largely based on the work of Arora and Zhang [4].

The majority of the separators currently used in batteries were typically developed as spin-offs of existing technologies. They were usually not developed specifically for those batteries and thus were not completely optimized for battery systems in which they are currently used. One positive result of adapting existing technologies is that the separators are produced in high volumes at relatively low cost. The availability of low-cost separators is an important consideration in the commercialization of batteries, since the battery industry traditionally operates with thin profit margins and relatively small research budgets.

Introduction and Scope

The battery industry has seen enormous growth over the past few years in portable, rechargeable battery packs. The majority of this surge can be attributed to the revolution in the use of multimedia in mobile phones and personal digital assistants (PDAs), as well in laptop computers and other wireless electronics devices. The introduction of vehicles implementing rechargeable batteries has increased the demand for batteries by several-fold. Batteries remained the mainstream source of power for systems ranging from mobile phones and personal digital assistants (PDAs) to electric and hybrid electric vehicles. The world market for batteries was approximately $41 billion in 2000, which included $16.2 billion primary and $24.9 billion secondary cells [20]. In 2010, the global demand was placed at $71 billion.

The Freedonia group estimates [21, 22] the aggregate US demand for primary and secondary batteries to be $16.8 billion in 2012 and that China will surpass the USA as the largest market with an estimated average annual growth at 7%. This growth will be driven by strong demand for battery-powered electronic devices like digital cameras and 3G wireless phones, increasing production of electrical and electronic equipment and the expansion in the automotive sector in the near future. The secondary battery demand has outpaced the primary battery market gains benefiting from strong growth in the use of high-drain portable electronic devices.

The tremendous progress in lithium-ion cells is clearly visible with as much as a threefold increase in the volumetric and gravimetric energy storage capability for both 18650 and prismatic cells between their initial introduction in the early 1990s and 2010. In last few years the lithium-ion production has expanded in South Korea (Samsung SDI, LG Chemical, etc.), China (ATL BYD, B&K, and Lishen, among others), and Japan. Several Japanese (Sanyo, Sony, MBI, NEC, etc.) and Korean (LG Chemical) manufacturers have also moved their manufacturing plants to China [23]. Japan, which controlled 94% of the global rechargeable battery market in 2000, has seen its market share drop to less than 50% of the global market [22, 24, 25]. The continued growth in lithium-ion battery market has led to a strong demand for battery separators. All the major separator manufacturers (Celgard, Asahi, and Tonen) have continually increased their capacity since 2003 [24, 26–29]. The industry has also witnessed the emergence of several new entrants [30–33].

There is not much information available on battery separator market in the literature. In 2009, it was estimated that about 50% of the rechargeable lithium battery market is the size of the components market, and separators constitute about 17% of this volume [22]. The Freedonia Group reported that the US demand for battery separators increased to $410 million in 2007 from $237 million in 1977, and $300 million in 2002, respectively [21, 22, 25].

The purpose of this chapter is to describe separators used in secondary batteries and characterization of their chemical, mechanical, and electrochemical properties, with particular emphasis on separators for lithium-ion batteries. The separator requirements, properties, and characterization techniques are described with respect to lithium-ion batteries. Despite the widespread use of separators, a need still exists for improving the performance, increasing its life, and extending the operating range.

Separator and Batteries

Batteries are built in many different shapes and configurations – button, flat, prismatic (rectangular), and cylindrical (AA, AAA, C, D, 18650 etc.). The cell components (including separators) are designed to accommodate a particular cell shape and design. The separators are either stacked between the electrodes or wound together with electrodes to form jellyrolls as shown in Fig. 1. Stacked cells are generally held together by pressure from the cell container. The lithium-ion gel polymer stacked cells are prepared by bonding/laminating layers of electrodes and separators together. The separator properties should not change significantly during the bonding process. In some cases, the separators are coated to help in bonding process and reduce the interfacial resistance [34, 35].

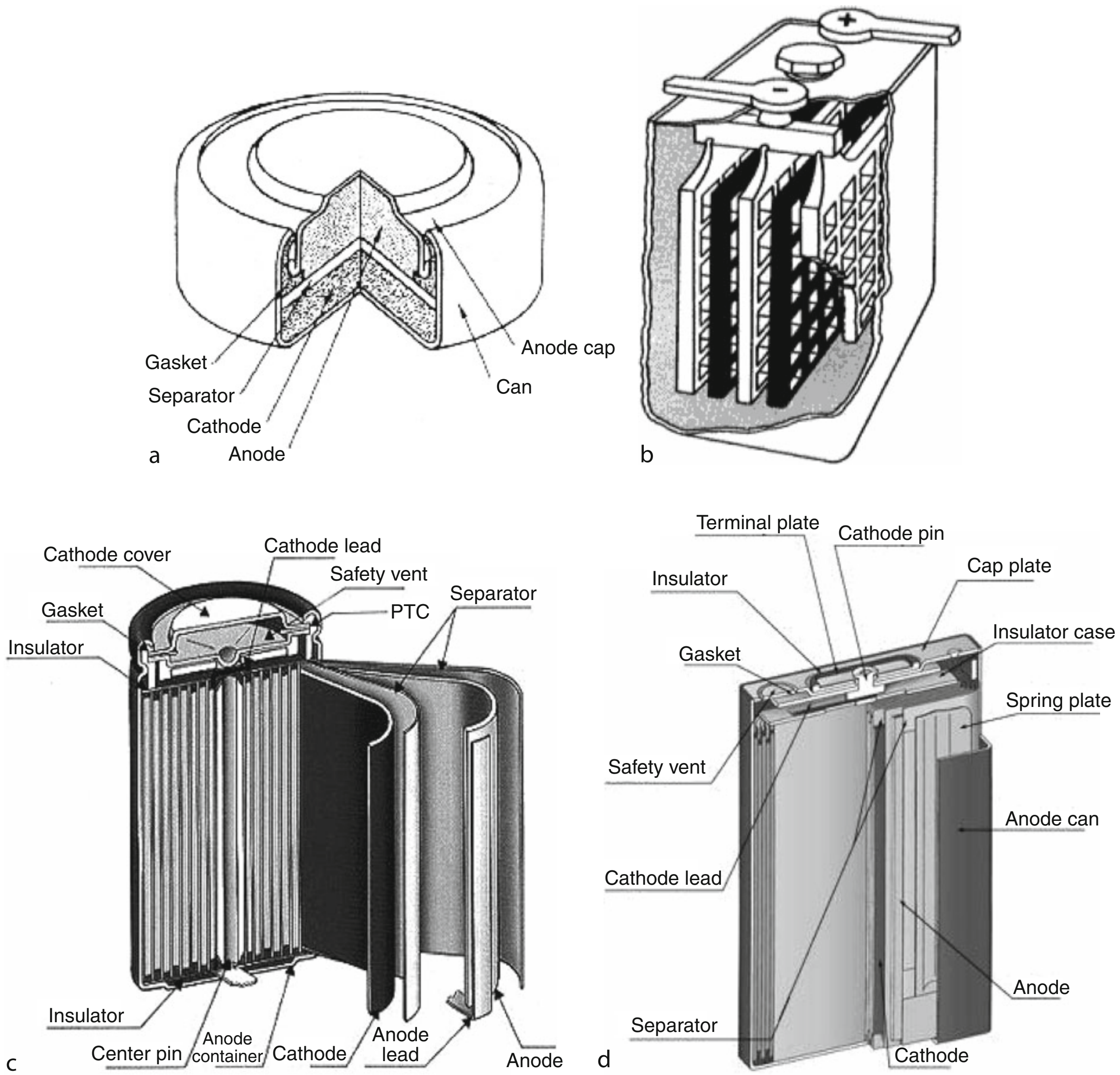
In the conventional way of making spirally wound cells, two layers of separators are wound along with the positive and negative electrodes, resulting in separator/negative/separator/positive configuration. They are wound as tightly as possible to ensure good interfacial contact. This requires the separators to be strong to prevent any contact between the electrodes through the separator. The separator also must not yield and reduce in width, or else the electrodes may contact each other. Once wound, the jellyroll is inserted into a can and filled with electrolyte. The separator must be wetted quickly by the electrolyte to reduce the electrolyte filling time. A header (cap) is then crimped onto the cell to cover the can top. In some prismatic cells, the electrode stack is pressed at high temperatures and pressures and then inserted into thin prismatic (rectangular) cans. A typical 18650 lithium-ion cell uses around 0.07–0.09 m^2 of separator, which is approximately 4–5% of the total cell weight [36].

A number of factors must be considered in selecting the best separator for a particular battery and application. The characteristics of each available separator must be weighed against the requirements and one selected that best fulfills these needs. A wide variety of properties are required of separators used in batteries. These include:

- Electronic insulator
- Minimal electrolyte (ionic) resistance
- Mechanical and dimensional stability
- Sufficient mechanical strength to allow easy handling
- Chemical resistance to degradation by electrolyte, impurities, and electrode reactants and products
- Effectiveness in preventing migration of particles between the two electrodes
- Readily wetted by electrolyte
- Uniformity in properties such as thickness, resistance, etc.

The above list presents a broad spectrum of requirements for separators in batteries. The order of importance of the various criteria varies, depending on the battery applications. In many applications, a compromise in requirements for the separator must be made to optimize performance, safety, cost etc. For example, batteries that are characterized by small internal resistance and consume little power require separators that are highly porous and thin; but the need for adequate physical strength may require that they be thick.

In addition to the above general requirements each battery type may have other specific requirements essential for good performance and/or safety. One typical example is that the separator used in sealed Nickel Metal Hydride (NiMH) batteries should be permeable to gas molecules for overcharge protection.



Rechargeable Batteries, Separators for. Figure 1
Typical battery configurations (**a**) button cell, (**b**) stack lead acid, (**c**) spiral wound cylindrical lithium-ion; (**d**) spiral wound prismatic lithium-ion

Separator Classification

Separators for batteries can be divided into different types, depending on their physical and chemical characteristics. They can be molded, woven, nonwoven, microporous, bonded, papers, or laminates. In recent years, there has been a trend to develop solid and gelled electrolytes that combine the electrolyte and separator into a single component.

In most batteries, the separators are either made of nonwoven fabrics or microporous polymeric films. Batteries that operate near ambient temperatures usually use separators fabricated from organic materials such as cellulosic papers, polymers, and other fabrics, as well as inorganic materials such as asbestos, glass-wool, and SiO_2. In alkaline batteries, the separators used are either regenerated cellulose or microporous polymer films. The lithium batteries with organic electrolytes mostly use microporous polymer films.

For the sake of discussion, the separators have been divided into six types – microporous films, nonwovens, ion-exchange membranes, supported liquid membranes, solid polymer electrolytes, and solid ion conductors. A brief description of each type of separator and their application in batteries are discussed below.

Microporous Separators

Separators are fabricated from a variety of inorganic, organic, and naturally occurring materials and generally contain pores that are greater than 50–100 Å in diameter. Materials such as nonwoven fibers (e.g., nylon, cotton, polyesters, glass), polymer films (e.g., polyethylene (PE), polypropylene (PP), polytetrafluoroethylene (PTFE), polyvinyl chloride (PVC)), and naturally occurring substances (e.g., rubber, asbestos, wood) have been used as microporous separators in batteries that operate at ambient and low temperatures (−40°C to 100°C). The microporous polyolefins (PP, PE, or laminates of PP and PE) are widely used in lithium-based nonaqueous batteries. More recently, other polymers have been employed, especially to enhance the window of temperature across which the polymer remains stable.

Nonwovens

Nonwovens are textile products that are manufactured directly from fibers. They are defined as a manufactured sheet, web, or matt of directionally or randomly oriented fibers, bonded by friction, and/or cohesion, and/or adhesion excluding paper and products which are woven, tufted, stitch-bonded incorporating binding yarns or filaments, or felted by wet milling whether or not needed. The fibers may be of natural or man-made origin. They may be staple or continuous filaments or maybe formed in situ [37].

The macroporous fibrous matrix is either dry laid, meltblown, or wet laid. The wet laid process is very similar to papermaking process. The fibers are bonded together by chemical or thermal bonding. The meltblown process is a binderless process and there the polymer fiber web is extruded. Typical pore size of fibrous matrix varies from 1 to 100 μm.

Nonwovens are widely utilized as separators for several types of batteries. Light-weight, wet laid nonwovens made from cellulosic, polyvinyl alcohol, and other fibers have achieved considerable success as separators for popular primary alkaline cells of various sizes. The key nonwoven attributes include consistently uniform basis weight, thickness, porosity, and resistance to degradation by electrolytes. Nonwovens are also successfully employed as separators in NiCd cells.

The materials used in nonwoven fabrics include a single polyolefin, or a combination of polyolefins, such as polyethylene (PE), polypropylene (PP), polyamide (PA), polytetrafluoroethylene (PTFE), polyvinylidine fluoride (PVdF), and polyvinyl chloride (PVC). Nonwoven fabrics have not been able to compete with microporous films in lithium-ion cells. This is primarily because of the inadequate pore-size structure and difficulty in making thin (<25 μm) nonwoven fabrics with acceptable physical properties. However, nonwoven separators have been used in button cells and bobbin cells when thicker separators and low discharge rates are acceptable.

Ion-Exchange Membranes

These membranes are generally fabricated from polymeric materials containing pores with diameters of less than 20 Å. The transport properties of ions in these membranes are characterized by strong interactions between the permeating species and the molecular structure of the polymer. This interaction is due to the presence of ion-exchange groups in the membrane, which allows the membrane to discriminate between permeating or migrating ions by virtue of their specific charge.

Radiation grafted membranes such as Permion® manufactured by RAI Research Corporation are ion-exchange membranes. Such membranes are used as battery separators in alkaline batteries. They are made from PE, PP, or Teflon-based films, which have excellent oxidation resistance and superior chemical resistance to alkali media. However, they are totally impervious to electrolyte flow, and therefore, have almost infinite resistance as a separator in this form. By using radiation grafting and cross-linking techniques, however, selected chemical species are grafted as pendant chains to the base structure of the linear polymer without altering the inert backbone. This modification imparts desirable hydrophilic properties to the films without materially impairing their excellent chemical resistance. This chapter provides a very limited discussion on ion-exchange membranes, as their application in batteries is very limited.

Supported Liquid Membranes

These types of separators consist of a solid matrix and a liquid phase, which is retained in the microporous structure by capillary forces. To be effective for batteries, the liquid in the microporous separator, which generally contains an organic phase, must be insoluble in the electrolyte, chemically stable, and still provide adequate ionic conductivity. Several types of polymers, such as polypropylene, polysulfone, polytetrafluoroethylene, and cellulose acetate, have been used for porous substrates for supported liquid membranes. The PVdF-coated polyolefin-based microporous membranes used in gel polymer lithium-ion battery fall into this category. Gel polymer electrolytes/membranes are only discussed briefly.

Polymer Electrolyte

Polymer electrolytes (e.g., poly(ethylene oxide), poly (propylene oxide)) have attracted considerable attention for batteries in recent years. These polymers form complexes with a variety of alkali metal salts to produce ionic conductors that serve as solid electrolytes. Its use in batteries is still limited due to poor electrode/electrolyte interface and poor room temperature ionic conductivity. Due to its rigid structure it can also serve as the separator. Polymer electrolytes are discussed briefly in the section Separators for Lithium-Ion Batteries.

Solid Ion Conductors

They serve as both separator and electrolyte. These are generally inorganic materials that are impervious barriers to gases and liquids. They allow one or more kinds of ions to migrate through their lattice when a potential gradient or a chemical gradient is present. These types of separators are beyond the scope of this article.

Separators for Lithium-Ion Batteries

All lithium-based batteries use nonaqueous electrolytes because of the reactivity of lithium in aqueous solution and because of electrolyte's stability at high voltage. The majority of these cells use microporous membranes made of polyolefins. In some cases, nonwovens made of polyolefins are either used alone or with microporous separators. This section will mainly focus on separators used in secondary lithium batteries.

Lithium secondary batteries can be classified into three types, a liquid-type battery using liquid electrolytes, a gel-type battery using gel electrolytes mixed with polymer and liquid, and a solid-type battery using polymer electrolytes. The types of separators used in different types of secondary lithium batteries are shown in Table 1. The liquid lithium-ion cell uses microporous polyolefin separators while the gel polymer lithium-ion cells either use PVdF separator (e.g., PLION® cells) or PVdF-coated microporous polyolefin separators. The PLION® cells use PVdF loaded with silica and plasticizer as separator. The microporous structure is formed by removing the plasticizer and then filling with liquid electrolyte. These are also characterized as plasticized electrolyte. In solid polymer lithium-ion cells, the solid electrolyte acts as both electrolyte and separator.

Sony's introduction of the rechargeable lithium-ion battery in the early 1990s precipitated a need for new

Rechargeable Batteries, Separators for. Table 1 Types of separators used in different types of secondary lithium batteries

Battery system	Type of separator	Composition
Lithium-ion (liquid electrolyte)	Microporous	Polyolefins (PE, PP, PP/PE/PP)
Lithium-ion gel polymer	Microporous	PVdF
	Microporous	Polyolefins (PE, PP, PP/PE/PP) coated with PVdF or other gelling agents
Lithium polymer (e.g., Li-V_6O_{13})	Polymer electrolyte	Polyethylene oxide with lithium salt

separators that provided good mechanical and electrical properties. Since then, separators have played a significant role in improving the performance as well as safety of lithium battery technology. In 2010, 90% of the total rechargeable battery markets for mobile communication devices use lithium-based batteries whereas the nickel metal hydride batteries continue to dominate the automotive sector and a significant factor in the camera market as well as alternate to the alkaline primary cell.

The microporous polyolefin separator has been used extensively in lithium-ion batteries, since it is difficult for most other conventional separator materials to satisfy the characteristics required in lithium-ion batteries. In lithium-ion batteries two layers of separators are sandwiched between positive and negative electrodes and then spirally wound together in cylindrical and prismatic configurations. The pores of the separator are filled with an ionically conductive liquid electrolyte.

Microporous polyolefin membranes (see Fig. 2) in current use are thin (<30 μm) and are made of polyethylene (PE), polypropylene (PP), or laminates [38] of polyethylene and polypropylene. Polyolefin materials are preferred because they provide excellent mechanical properties, chemical stability, and acceptable cost [39, 40]. They have been found to be compatible with the cell chemistry and can be cycled for several hundred cycles without significant degradation in chemical or physical properties.

Commercial membranes offer pore size in the range of 0.03–0.1 μm, and 30–50% porosity. The low melting point of PE enables their use as a thermal fuse. As the temperature approaches the melting point of the polymer, 135°C for PE and 165°C for PP, the porosity of the membrane is lost. The trilayer material (PP/PE/PP) [41] has been developed by Celgard® where a PP layer is designed to maintain the integrity of the film, while the low melting point of PE layer is intended to shutdown the cell if an over-temperature condition is reached [42]. Asahi Kasai's flat-film membrane "Hipore™" is available in thicknesses ranging from 20 μm to several hundred micrometers, and with highly uniform pore sizes ranging from 0.05 to 0.5 μm [43]. The major manufacturers of lithium-ion battery separators along with their typical products are listed in Table 2.

In recent years there has been a strong demand for higher-capacity lithium-ion cells because of the strong growth in portable electronics. One way to achieve higher capacity is by reducing the thickness of separators. At present, battery manufacturers routinely use separators 16 μm or thinner in higher-capacity (>2.6 Ah) cylindrical cells and 9 μm separators in lithium-ion gel polymer cells.

Separator Development

The process for making lithium-ion battery separators can be broadly divided into dry and wet processes. Both processes usually employ one or more orientation steps to impart porosity and/or increase tensile strength. Dry process involves melting a polyolefin resin, extruding it into a film, thermal annealing to increase the size and amount of lamellar crystallites, and precisely stretching to form tightly ordered micropores [44–48]. In this

Rechargeable Batteries, Separators for. Figure 2
Polyolefin separators used in lithium-ion batteries

Rechargeable Batteries, Separators for. Table 2 Major manufacturers of lithium-ion battery separators along with their typical products

Manufacturer	Structure	Composition	Process	Trade name
Asahi Kasai	Single layer	PE	Wet	HiPore
Celgard Inc	Single layer	PP, PE	Dry	Celgard
	Multilayer	PP/PE/PP	Dry	Celgard
	PVdF coated	PVdF, PP, PE, PP/PE/PP	Dry	Celgard
Entek Membranes	Single layer	PE	Wet	Teklon
Mitsui Chemical	Single layer	PE	Wet	
Nitto Denko	Single layer	PE	Wet	
DSM	Single layer	PE	Wet	Solupur
Tonen	Single layer	PE	Wet	Setela
Ube Industries	Multilayer	PP/PE/PP	Dry	U-Pore

process, a row lamellar crystal structure is generated in the polymer in the initial extrusion step. This nonporous structure is highly oriented as a result of extrusion and annealing conditions. The films are then stretched to form micropores. This microporous structure is continuous throughout the bulk interior of the membrane [49].

Polypropylene and polyethylene microporous films obtained by this method are available from Celgard, [44, 46, 50, 51] and Ube [52]. The dry process is technologically convenient and environmentally benign because no solvents are required. However, biaxial stretching has been met with limited success to date and, as a result, the pores are slit-like in shape, and the mechanical properties of films are anisotropic. The tensile strength in the lateral direction is relatively low.

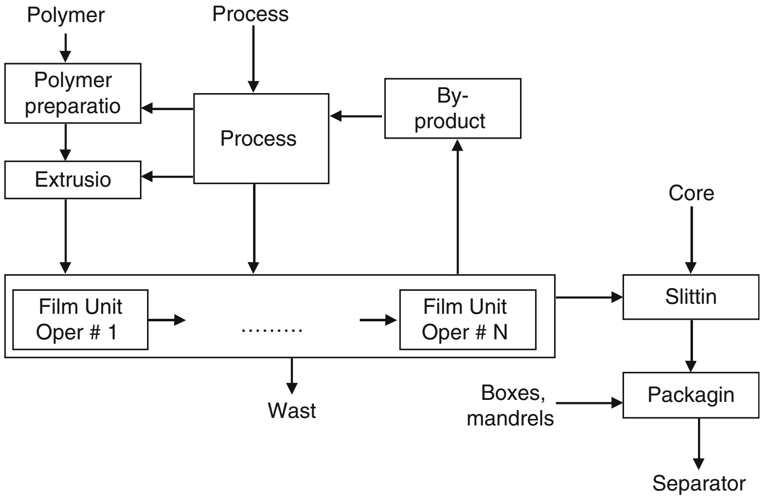
Wet process (phase inversion process) [53, 54] involves mixing of hydrocarbon liquid or some other low-molecular-weight substance generally with a polyolefin resin, heating and melting the mixture, extruding the melt into a sheet, orientating the sheet either in the machine direction (MD) or biaxially, and then extracting the liquid with a volatile solvent [55, 56]. Separators made by wet process are available from Asahi Kasei [57], Tonen [58–60], Mitsui Chemicals [57], and more recently from Polypore/ Membrana and Entek [31]. The structure and properties of the membranes can be changed by controlling the composition of the solutions and the evaporation or removal of solvents in the gelation and solidification processes. The separators made by wet process use ultrahigh-molecular-weight polyethylene (UHMWPE). The use of UHMWPE gives good mechanical properties as well as some degree of melt integrity.

Ihm et al. have given an overview of the wet process by preparing a separator with polymer blends of high-density polyethylene (HDPE) and ultrahigh-molecular-weight polyethylene (UHMWPE) [54]. They showed that the mechanical strength and drawing characteristics are influenced by the content and the molecular weight of the UHMWPE contained in a polyolefin blending solution. The manufacturing process of typical microporous film by dry and wet process is compared in Table 3.

A simplified flowchart for separator manufacturing process is shown in Fig. 3 [61]. The virgin polymer is prepared and mixed with processing aids (e.g., antioxidants, plasticizer, etc.) and then extruded. The extruded polymer then goes through different steps, which vary from process to process. For dry process, it can involve film annealing and stretching, while for wet process, it can involve solvent extraction and stretching. The finished film is then slit into required widths and packed into boxes and shipped to the battery manufacturers. With the advent of thinner separators, the film handling during manufacturing steps has become very important for the final quality of the film.

Rechargeable Batteries, Separators for. Table 3 Manufacturing process of typical microporous film

Process	Mechanism	Raw material	Properties	Typical membranes	Manufacturers
Dry process	Drawing	Polymer	Simple process Anisotropic film	PP, PE, PP/PE/PP	Celgard, Ube
Wet process	Phase separation	Polymer + Solvent	Isotropic film	PE	Asahi, Tonen
		Polymer + Solvent + Filler	Large pore size High porosity	PE	Asahi



Rechargeable Batteries, Separators for. Figure 3
Generalized process for lithium-ion separator manufacturing [61]. Each step of the separator manufacturing process has online detection systems to monitor the quality of the separator

Each step of the separator manufacturing process has online detection systems to monitor the quality of the separators.

Uniaxially oriented films generally have high strength in only one direction, whereas biaxially oriented films are relatively strong in both machine direction (MD) and transverse direction (TD). However, biaxial orientation often tends to introduce TD shrinkage. This shrinkage, at elevated temperatures, can allow electrodes to contact each other. The separator must have sufficient strength in the machine direction so that it does not decrease in width or break under the stress of winding. The typical requirement for the mechanical strength in a 25-μm separator is 1,000 kg/mm^2 [54].

The typical properties of some commercial microporous membranes are summarized in Table 4. Celgard 2730 and Celgard 2400 are single-layer PE and PP separators, respectively, while Celgard 2320 and 2325 are Trilayer separators of 20 and 25 μm thickness. Asahi and Tonen separators are single-layer PE separators made by wet process. Basic properties such as thickness, Gurley, porosity, melt temperature, and ionic resistivity are reported in Table 4. These properties are defined in a subsequent section.

Efforts have been made to find a new route for dry process using biaxial stretching techniques for preparing polypropylene microporous films, which may have submicron pore size and narrow size distribution, high permeability to gasses and liquids combined with good mechanical properties. The biaxially stretched polypropylene microporous films (Micpor®) were made by using nonporous polypropylene films of high beta-crystal content [62]. The porosity of these films can be as high as 30–40%, with an average pore size of

Rechargeable Batteries, Separators for. Table 4 Typical properties of some commercial microporous membranes

Separator/properties	Celgard 2730	Celgard 2400	Celgard 2320	Celgard 2325	Asahi Hipore	Tonen Setela
Structure	Single Layer	Single layer	Trilayer	Trilayer	Single layer	Single layer
Composition	PE	PP	PP/PE/PP	PP/PE/PP	PE	PE
Thickness (um)	20	25	20	25	25	25
Gurley (s)	22	24	20	23	21	26
Ionic resistivity[a] (Ω-cm^2)	2.23	2.55	1.36	1.85	2.66	2.56
Porosity (%)	43	40	42	42	40	41
Melt temperature (°C)	135	165	135/165	135/165	138	137

[a]In 1 M $LiPF_6$ EC:EMC (30:70 by volume)

approximately 0.05 μm. The pores on the surface were almost circular in shape compared to slit-like pores observed in uniaxial stretched samples and exhibited high permeability to fluids with good mechanical properties and almost circular pore shape with narrow pore-size distribution [63–65].

The PP/PE bilayer [38] and PP/PE/PP trilayer separators were developed by Celgard. Multilayer separators offer advantages of strength and combine the lower melting temperature of PE with the high-temperature strength of PP. Nitto Denko has also patented a single-layer separator made from a blend of PE/PP by the dry stretch process [66]. According to the patent, the separator has microporous regions of PE and PP. On heating in an oven, the impedance of the separator increases near the melting point of PE and the impedance remains high until beyond the melting point of PP. However, battery performance data have not been presented.

Microporous polyethylene separator material composed of a combination of randomly oriented thick and thin fibrils of ultrahigh-molecular-weight polyethylene (UHMWPE), Solupur®, manufactured by DSM Solutech, is also an interesting separator material for lithium-ion batteries. Solupur® is fabricated in standard grades with base weights ranging from 7 to 16 g/m^2 and mean pore size ranging from 0.1 to 2.0 μm and a porosity of 80–90% [67]. Ooms et al. carried out a study on a series of DSM Solupur materials with different permeability. Rate capability and cycling tests of these materials were compared with commercial available separators in CR2320-type coin cells. Solupur® materials showed low tortuosity, high strength and puncture resistance, excellent wettability, good high rate capability and low-temperature performance because of its high porosity and UHMWPE structure [68].

Nitto Denko has also developed a battery separator made by a wet process that had high puncture strength and high heat rupture resistance [69]. They used a polyolefin resin with a high-molecular-weight rubber as its main component material and cross-linked through oxidation in air. The melt rupture temperature, as measured by thermomechanical analysis, was over 200°C in this material. They also tried cross-linking ultrahigh-molecular-weight polyethylene with electron-beam and ultraviolet irradiation, but this had the side effect of causing deterioration in the polyolefin including rupture of the main chains and therefore resulted in reduced strength.

ENTEK Membranes LLC has developed Teklon™ – a highly porous, ultrahigh-molecular-weight polyethylene separator for lithium-ion batteries. At the writing of this publication, the separator is available in small quantities. Pekala et al. characterized Celgard™, Setela™, and Teklon™ separators in terms of their physical, mechanical, and electrical properties [70].

Celgard's separators are the best-characterized battery separators in literature as they have been widely used in numerous battery systems. Bierenbam et al. [55] has described the process, physical, and chemical properties, and end-use applications. Fleming and Taskier [71] described the use of Celgard microporous membranes as battery separators. Hoffman et al. [72]

presented a comparison of PP and PE Celgard microporous materials. Callahan discussed a number of novel uses of Celgard membranes. Callahan and coworkers [73] also characterized Celgard membranes by SEM image analysis, mercury porosimetry, air permeability, and electrical resistivity, and later characterize the puncture strength and temperature/impedance data for Celgard membranes [38]. Spotnitz et al. reported short-circuit behavior in simulated, spirally wound cells, as well as impedance/temperature behavior and thermomechanical properties [42]. Yu [74] found that a trilayer structure of PP/PE/PP Celgard™ microporous membranes provided exceptional puncture strength.

Nonwoven materials such as cellulosic fibers have never been successfully used in lithium batteries. This lack of interest is related to the hygroscopic nature of cellulosic papers and films, their tendency to degrade in contact with lithium metal, and their susceptibility to pinhole formation at thickness of less than 100 µm.

Asahi Chemical Industry carried out an exploratory investigation to determine the requirements for cellulose-based separators for lithium-ion batteries [75]. In an attempt to obtain an acceptable balance of lithium-ion conductivity, mechanical strength, and resistance to pinhole formation, they fabricated a composite separator (39–85 µm) that consists of fibrilliform cellulosic fibers (diameter 0.5–5.0 µm) embedded in a microporous cellulosic (pore diameter: 10–200 nm) film. The fibers are intended to reduce the possibility of separator meltdown under exposure to heat generated by overcharging or internal short circuiting. The resistance of these films was equal to, or lower than, the conventional polyolefin-based microporous separators. The long-term cycling performance was also very comparable.

Pasquier et al. [76] used paper-based separators in flat pouch-type lithium-ion batteries and compared the performance with cells made with Celgard-type polyolefin-based separators. The paper separators had good wetting properties, good mechanical properties, but did not provide the shutdown effect essential for large lithium-ion batteries. Their resistance was similar to polyolefin separators and when all water traces were removed from paper, their cycling performance was similar to Celgard separators. The paper-based separators can be used in small flat pouch-type cells where high strength and shutdown behavior is not required. For larger spherically wound cells, which require strong separators with shutdown feature, paper-based separators cannot be used.

Recently Degussa announced that they have developed Separion® separators for lithium batteries by combining the characteristics of flexible polymeric separators with the advantages of chemical and thermally resistant and hydrophilic ceramic materials. Separion® is produced in a continuous coating process. Ceramic materials, e.g., alumina, silica, and/or zirconia, are slip coated and hardened onto a support [77, 78]. According to Degussa, Separion separators have an excellent high temperature stability, superior chemical resistance, and good wettability, especially at low temperatures. They tested the performance and safety behavior of Separion separator in 18,650 cells and found the performance to be comparable to polyolefin-based separators [79].

The potential use of polymeric ion-exchange membranes in the next generation single-ion secondary lithium polymer batteries was shown by Sachan et al. [80, 81]. Conductivities exceeding 10^{-4} S/cm with transference numbers of unity were achieved for Nafion converted to Li^+ salt form. However, little work has continued to introduce ion-exchange membranes in lithium batteries. The cost associated with the manufacturing of such membranes has been a significant barrier to their commercial viability.

To obtain a thin (less than 15 µm) separator for lithium batteries, Optodot has taken a different approach of high-speed coating of a metal oxide sol gel coating on a smooth surface followed by a delamination step to provide the free-standing separator. Using this approach, separator with thicknesses from 6 to 11 µm was made on large-scale production coating equipment [82]. They found that the sol gel separators with a thickness in the middle of this range of 8–9 µm have the preferred combination of thinness and strength. The metal oxide sol gel coating is water based with no organic solvents present. The coating formulations include a polymer and a surfactant. The polymer provides improved coating rheology, mechanical strength, and other properties. The surfactant provides improved wetting properties on the substrate. The films prepared were around 11 µm thick, with 45% porosity, completely wettable in nonaqueous electrolyte, and melt temperature greater than 180°C. While these films are relatively thin and should help in increasing the capacity, they may not be strong enough for use in tightly wound cells. Moreover, the shutdown temperature of the separator seems to be very high and thus not suitable for existing lithium-ion chemistries.

Gineste et al. carried out the grafting of hydrophilic monomers onto PP or PE separators to improve the wettability of separators used in secondary lithium batteries with a lower content of wetting agents [83, 84]. They used a PP film (Celgard 2505) of 50 μm thickness after irradiating in air by electron beams with a dose ranging from 0.5 to 4 Mrad. The irradiated film was grafted by a monofunctional monomer (acrylic acid, AA), in the presence of difunctional cross-linking agent (diethyleneglycol dimethacrylate, DEGDM). The separators start losing mechanical properties, when the grafting ratio is higher than 50%.

Separator Requirements

In lithium-based cells, the essential function of battery separator is to prevent electronic contact, while enabling ionic transport, between the positive and negative electrodes. It should be usable on high-speed winding machines and possess good shutdown properties. The most commonly used separators for primary lithium batteries are microporous polypropylene membranes. Microporous polyethylene and laminates of polypropylene and polyethylene are widely used in lithium-ion batteries [85]. These materials are chemically and electrochemically stable in secondary lithium batteries. A key requirement for the separators for lithium primary batteries is that their pore size be small enough to prevent dendritic lithium penetration through them.

The general requirements [86] for lithium-ion battery separators are given below:

1. *Thickness* – The lithium-ion cells used in consumer applications use thin microporous separators (<25 μm). The separators being developed for EV/HEV applications will require thicker (∼40 μm) separators. The thicker the separator, the greater the mechanical strength and the lower the probability of punctures during cell assembly, but reduce the amount of active materials that can be placed in the same size can. The thinner separators lower the internal resistance, take up less space, and permit the use of longer electrodes for increased capacity and rate capability.
2. *Permeability* – The separators should not limit the electrical performance of the battery under normal conditions. Typically the presence of separator increases the effective resistivity of the electrolyte by a factor of 6–7. The ratio of the resistivity of the separator filled with electrolyte divided by the resistivity of the electrolyte itself is called MacMullin number. The rate capability of the battery is inversely proportional to the MacMullin number; however, it has been shown that increase in resistance across the electrode material has a larger impact on the cell performance for comparable changes made to the separator [87].
3. *Gurley* (*air permeability*) – Air permeability is proportional to electrical resistivity, for a given separator morphology. It is used in place of electrical resistance (ER) measurements once the relationship between Gurley and ER is established. The separator should have low Gurley values for good electrical performance.
4. *Porosity* – It is implicit in the permeability requirement; typically lithium-ion battery separators have a porosity of 40%. Control of porosity is very important for battery separators. Specification of percent porosity is commonly an integral part of separator acceptance criteria. The porosity of separators used in alkaline zinc MnO_2 cells is typically around 80–90%.
5. *Wettability* – The separators should wet out quickly and completely in typical battery electrolytes. The lack of wetting results in localized spots of high resistance.
6. *Electrolyte absorption and retention* – A separator should be able to absorb and retain electrolyte. Electrolyte absorption is needed for ion transport. The microporous membranes usually do not swell on electrolyte absorption.
7. *Chemical stability* – The separators should be stable in battery for a long period of time. It should be inert to both strong reducing and strong oxidizing conditions and should not degrade or loose mechanical strength or produce impurities, which can interfere with the function of the battery. The separator must be able to withstand the strong

oxidizing positive electrode and the corrosive nature of the electrolyte at temperatures as high as 75°C. The greater the oxidation resistance, the longer the separator will survive in a cell. Polyolefins (e.g., polypropylene, polyethylene) exhibit high resistance to most of the conventional chemicals, good mechanical properties, and moderate temperature range for application making it one of the ideal polymers for lithium-ion battery separators. Polypropylene separators exhibit better mechanical properties and minimal oxidation when in contact with the positive electrode in a lithium-ion cell. Thus, the performance of trilayer (PP/PE/PP) separators with PP as the outside layer and PE as inner layer are superior.

8. *Dimensional stability* – The separator should lay flat and should not curl at the edges when unrolled as this can greatly complicate cell assembly. The separator should also not shrink when exposed to electrolyte. The cell winding should not affect the porous structure in any adverse way.
9. *Puncture strength* – The separators used in wound cells require high puncture strength to avoid penetration of electrode material through the separator. If particulate material from the electrodes penetrates the separator, an electrical short will result and the battery will be rejected. The separators used in lithium-ion batteries require more strength then the one used in lithium primary batteries. The primary lithium batteries have only one rough electrode and thus it requires less strength. As empirically observed, for most applications, the puncture strength should be at least 400 g/mil for separators used in lithium-ion cells. Mix penetration strength is a better measure of separator strength in a battery compared to puncture strength.
10. *Mix penetration strength* – The susceptibility of separators to particle penetration is characterized by mix penetration strength [45]. During the winding of the spiral wrap construction considerable mechanical pressure is applied to the cathode–separator–anode interface. Any loose particle could be forced through the separator and short the cell. The mix penetration strength should be at least 100 kgf/mil for separators used in lithium-ion cells.
11. *Thermal stability* – Lithium-ion batteries can be poisoned by water and so materials going into the cell are typically dried at 80°C under vacuum. Under these conditions, the separator must not shrink significantly and definitely must not wrinkle. Each battery manufacturer has specific drying procedures. The requirement of less than 5% shrinkage after 60 min at 90°C (in vacuum) in both MD and TD direction is a reasonable generalization.
12. *Pore size* – A key requirement of separators for lithium batteries is that their pores be small enough to prevent dendritic lithium penetration through them. Membranes with submicron pore sizes have proven adequate for lithium batteries.
13. *Tensile strength* – The separator is wound with the electrodes under tension. The separator must not elongate significantly under tension in order to avoid contraction of the width. A tensile strength specification is sometimes given, but the key parameter is Young's Modulus in the machine direction. Since Young's Modulus is difficult to measure, 2% offset yield is a good measure; less than 2% offset at 1,000 psi is acceptable for most winding machines.
14. *Camber* – Ideally, when a strip of separator is laid out, the separator should be straight and not bow or skew. In practice, however, some camber is often observed. If sufficiently extreme, this can cause misalignment between the electrodes and separator. Camber can be measured by laying the separator flat on a table parallel with a straight meter stick. The camber should be less than 0.2 mm/m of separator.
15. *Shutdown* – Lithium-ion battery separators provide some margin of protection against short circuit and overcharge in lithium-ion cells. The separators exhibit a large increase in impedance at temperature about 130°C that effectively stops ionic transport between the electrodes [88, 89]. The greater the mechanical integrity of the separator above 130°C, the greater the margin of safety the separator can provide. If the separator looses mechanical integrity, then the electrodes can come into direct contact, react chemically, and result in thermal runaway. The shutdown behavior of a separator can be characterized by heating the

separator (saturated with electrolyte) to high temperatures and simultaneously monitoring the electrical resistance of the separator [42, 89].

16. *High temperature stability* – A separator might provide an extra margin of safety if it can prevent the electrodes from contacting one another at high temperatures. Separators with good mechanical integrity at high temperatures can provide a greater margin of safety for lithium-ion cells. Thermal Mechanical Analysis (TMA) can be used to characterize the high temperature stability of separators. Utilizing TMA, the separator is held under constant load and the degree of elongation versus temperature is measured; the temperature at which the separator loses mechanical integrity, the elongation increases dramatically.
17. *Electrode interface* – The separator should form a good interface with the electrodes to provide sufficient electrolyte flow.

In addition to the above properties, the separator must be essentially free of any type of defects (pinholes, gels, wrinkles, contaminants, etc.). All of the above properties have to be optimized before a membrane qualifies as a separator for a lithium-ion battery. The general requirements for lithium-ion battery separators are also summarized in Table 5.

Separator Properties/Characterization

Separators are characterized by structural and functional properties; the former describes what they are and the latter how they perform. The structural properties include chemical (molecular) and microcrystalline nature, thickness, pore size, pore-size distribution, porosity, and various chemical and physical properties such as chemical stability, and electrolyte uptake. The functional properties of interest are electrical resistivity, permeability, and transport number. It is useful to characterize separator materials in terms of their structural and functional properties, and to establish a correlation of these properties with their performance in batteries. A variety of techniques are used to evaluate separators. Some of these techniques are discussed in this section.

Gurley Number Separator permeability is usually characterized by air permeability. The Gurley number expresses the time required for a specific amount of air to pass through a specific area of separator under a specific pressure. The standard test method is described in ASTM-D726 (B).

The Gurley number is used to characterize separators because the measurement is accurate and easy to make, and deviations from specific values are a good indication of problems. Air permeability (Gurley) is proportional to electrical resistance (ER), for a given separator morphology [90]. Gurley can be used in place of ER measurements once the relationship between Gurley and ER is established. A lower Gurley value means higher porosity, lower tortuosity, and accordingly lower ER.

Electrical Resistance The measurement of separator resistance is very important to the art of battery manufacture because of the influence the separator has on electrical performance. Electrical resistance is a more comprehensive measure of permeability then the Gurley number, in that the measurement is carried out in the actual electrolyte solution. The ionic resistivity of the porous membrane is essentially the resistivity of the electrolyte that is embedded in the pores of the separator. Typically, a microporous separator, immersed in an electrolyte, has an electrical resistivity about six to seven times that of a comparable volume of electrolyte, which it displaces. It is a function of the membrane's porosity, the tortuosity, the resistivity of the electrolyte, the thickness of the membrane, and the extent to which the electrolyte wets the pores of the membrane [83]. The electrical resistance of the separator is the true performance indicator of the cell. It describes a predictable voltage loss within the cell during discharge and allows one to estimate rate limitations.

Classical techniques for measuring electrical resistivity of microporous separators have been described by Falk and Salkind [5] and Robinson and Walker [42]. The resistivity of an electrolyte is more accurately determined by AC methods since DC can polarize the electrodes and cause electrolysis of the solution. Modern AC impedance measuring systems allow rapid measurements of cell resistance over a wide range or frequencies from which resistance can be calculated free of capacitance effects. Compared to the DC techniques, the equipment required and the theory

Rechargeable Batteries, Separators for. Table 5 General requirements for lithium-ion battery separator [86]

Parameter	Goal
Thickness[a,b] (μm)	<25
Electrical resistance (MacMullin No.[c], dimensionless)	<8
Electrical resistance (ohms-cm^2)	<2 Ω-cm^2
Gurley[d] (s)	~25 ml
Pore size[e] (μm)	<1
Porosity (%)	~40
Puncture strength[f] (g)	>300 g/mil
Mix penetration strength (kgf)	>100 kgf/mil
Shrinkage[g] (%)	<5% in both MD and TD
Tensile strength[h]	<2% offset at 1,000 psi
Shutdown temperature (°C)	~130
High temperature melt integrity (°C)	>150
Wettability	Complete wet out in typical battery electrolytes
Chemical stability	Stable in battery for long period of time
Dimensional stability	Separator should lay flat; be stable in electrolyte
Skew	<0.2 mm/m

[a] ASTM D5947-96, "Standard test methods for physical dimensions of solid plastics specimens," ASTM International.

[b] ASTM D2103, "Standard specification for polyethylene film and sheeting," ASTM International.

[c] Caldwell DL, Poush KA (1984) US Patent 4,464,238

[d] ASTM D726, "Standard test methods for identification of fibers in textiles," ASTM International.

[e] ASTM E128-99, "Standard test method for maximum pore diameter and permeability of rigid porous filters for laboratory use," ASTM International.

[f] ASTM D3763, "Standard test method for high-speed puncture properties of plastics using load and displacement sensors," ASTM International.

[g] ASTM D1204, "Standard test methods for linear dimensional changes of nonrigid hermoplastic sheeting or film at elevated temperatures," ASTM International.

[h] ASTM D882, "Standard test method for tensile properties of thin plastic sheeting," ASTM International.

necessary to interpret the AC techniques are more complex; however, AC measurements yield information about long-range migration of ions and polarization phenomena occurring within the cell. In an AC measurement, a sinusoidal voltage is applied to a cell, and the sinusoidal current passing through the cell as a result of this perturbation is determined. A four-electrode cell is usually used for resistivity measurements. The outer two electrodes serve to apply a sinusoidal potential, and the resulting current passing through the inner two electrodes is measured. This technique is employed to avoid the complications arising from a nonuniform potential field near the outer two electrodes. An excellent review of experimental techniques for measuring electrical resistivity in aqueous solution is available [91, 92].

The separator resistance is usually characterized by cutting small pieces of separators from the finished material and then placing them between two blocking electrodes. The separators are completely saturated with the electrolyte. The resistance (Ω) of the separator is measured at a certain frequency by AC impedance techniques. The frequency is chosen so that the separator impedance is equal to the separator resistance. In order to reduce the measurement error, it is best to do multiple measurements by adding extra layers.

The average resistance of single layer is determined from multiple measurements. The specific resistivity, ρ_s (Ω-cm), of the separator saturated with electrolyte is given by

$$\rho_s = \frac{R_s A}{l} \tag{1}$$

where R_s is the measured resistance of separator in Ω, A is the electrode area in square centimeters and l is the thickness of membrane in centimeters. Similarly, the specific resistivity of the electrolyte, ρ_e (Ω-cm), is given by

$$\rho_e = \frac{R_e A}{l} \tag{2}$$

where R_e is the measured resistance of electrolyte in ohms. The ratio of the resistivity of a separator membrane to that of the electrolyte is called MacMullin number, N_m, which can be used to predict the influence of the separator on battery performance [93].

$$N_m = \frac{\rho_s}{\rho_e} = \frac{\tau^2}{\varepsilon} \tag{3}$$

where τ is the tortuosity and ε is the porosity of the separator. The MacMullin number describes the relative contribution of a separator to cell resistance. It is almost independent of electrolyte used and also factors out the thickness of the material. It assumes that the separator wets completely in the electrolyte used for the test. From Eqs. 1 and 3 the electrical resistance of a microporous membrane is given by the following [5, 94]:

$$R_m = \rho_e\left(\frac{\tau^2 l}{\varepsilon A}\right) \tag{4}$$

It has been shown for Celgard membranes that the membrane resistance can be related to the Gurley number by Eq. 5. where R_m is the membrane resistance (Ω), A is the membrane area (cm^2), ρ_e is the specific electrolyte resistance (Ω-cm), t_{gur} is the Gurley number (10 cm^3 air, 2.3 mmHg), d is the pore size, and 5.18×10^{-3} a scaling factor [90].

$$R_m A = \frac{\rho_e}{5.18 \times 10^{-3}} t_{\text{gur}} d \tag{5}$$

The usual procedure for characterizing battery separators is to cut several test samples from the finished material. Thus, only a small portion of the separator is actually examined. Ionov et al. has proposed an alternative technique to measure the resistance of a separator over a large separator area [95]. In this technique the separator material is passed through an electrolyte bath between electrical resistance measuring transducers. The set of transducers installed in the bath transverse to the moving sheet of separator material examines the whole surface of the material. If the production process ensures good uniformity in the physicochemical properties of the separator material over the whole surface, the transducer outputs will be close to one another. A nonuniform separator will cause significant deviations from the average value at various sections of the material examined. In this case, the sections having lower or higher resistance compared with the average value should be regarded as flawed.

Porosity The porosity is important for high permeability and also for providing a reservoir of electrolyte in the cell. A higher and uniform porosity is desirable for unhindered ionic current flow. Nonuniform porosity leads to nonuniform current density and can further lead to reduced activity of the electrodes. Cell failure can result, if during discharge, some areas of the electrodes work harder than other.

Porosity of a separator is defined as the ratio of void volume to apparent geometric volume. It is usually calculated (Eq. 6) from the skeletal density, basis weight, and dimensions of the material and so may not reflect the accessible porosity of the material.

$$\text{Porosity}(\%) = 1 - \frac{\left(\frac{\text{Sample weight}}{\text{Sample volume}}\right)}{\text{Polymer density}} \times 100\% \tag{6}$$

The standard test method is described in ASTM D-2873. The actual or accessible porosity can also be determined by the weight of liquid (e.g., Hexadecane) absorbed in the pores of the separator. In this method, the separator weight is measured before and after dipping in Hexadecane solvent and the porosity is calculated (Eq. 7) by assuming that volume occupied by Hexadecane is equal to the porous volume of the separator.

$$\text{Porosity}(\%) = \frac{\text{Volume occupied by Hexadecane}}{\text{Volume of Polymer Volume occupied by Hexadecane}} \times 100 \tag{7}$$

Tortuosity Tortuosity is the ratio of mean effective capillary length to separator thickness. The tortuosity factor τ of a separator can be expressed by

$$\tau = \frac{l_s}{d} \tag{8}$$

where l_s is the ion path through the separator and d is the thickness of the separating layer.

Tortuosity is a long-range property of a porous medium, which qualitatively describes the average pore conductivity of the solid. It is usual to define τ by electrical conductivity measurements. With knowledge of the specific resistance of the electrolyte and from a measurement of the sample membrane resistance, thickness, area, and porosity, the membrane tortuosity can be calculated from Eq. 3.

This parameter is widely used to describe the ionic transport by providing information on the effect of pore blockage. A tortuosity factor $\tau = 1$, therefore, describes an ideal porous body with cylindrical and parallel pores, whereas values of $\tau > 1$ refer to more hindered systems. Higher tortuosity is good for dendrite resistance but can lead to higher separator resistance.

Pore Size and Pore-Size Distribution For any battery applications, the separator should have uniform pore distribution to avoid performance losses arising from nonuniform current densities. The submicron pore dimensions are critical for preventing internal shorts between the anode and the cathode of the lithium-ion cell, particularly since these separators tend to be as thin as 25 μm or less. This feature will be increasingly important as battery manufacturers continue to increase the cell capacity with thinner separators. The pore structure is usually influenced by polymer composition, and stretching conditions, such as drawing temperature, drawing speed, and draw ratio. In the wet process, the separators produced by the process of drawing after extraction (as claimed by Asahi Chemical and Mitsui Chemical) are found to have much larger pore size (0.24–0.34 μm) and wider pore-size distribution than those produced by the process of extraction (0.1–0.13 μm) after drawing (as claimed by Tonen) [54].

The testing of battery separators and control of their pore characteristics are important requirements for proper functioning of batteries. Mercury porosimetry has been historically used to characterize the separators in terms of percentage porosity, mean pore size, and pore-size distribution [96]. In this method, the size and volume of pores in a material are measured by determining the quantity of mercury, which can be forced into the pores at increasing pressure. Mercury does not wet most materials and a force must be applied to overcome the surface tension forces opposing entry into the pores.

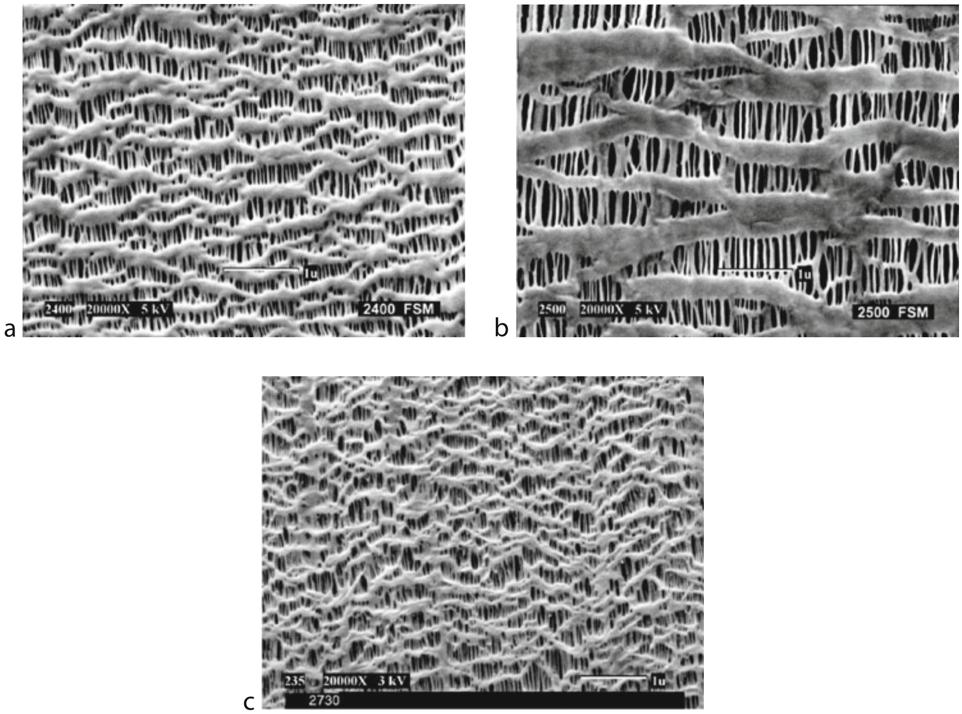
The hydrophobic (e.g., polyolefins) separators are also characterized with Aquapore (nonmercury porosimetry) technique, where water is used in place of mercury. This is a very useful technique for characterizing polyolefin-based separators used in lithium batteries [97]. Porosimetry gives pore volume, surface area, mean pore diameter, and pore-size distribution. In a typical experiment, the sample is placed in the instrument and evacuated. As the pressure increases, the quantity of water forced into the pores increases in proportion to the differential pore volume, the size of the pores corresponding to the instantaneous pressure. Thus, increasing the pressure on a membrane having a given pore-size distribution results in a unique volume versus pressure or pore diameter curve. The pressure required for intrusion of water into a pore of diameter D is given by following equation

$$D = \frac{4\gamma \cos\theta}{p} \tag{9}$$

where D is the diameter of the pore assuming the pore to be cylindrical, p is the differential pressure, γ is the surface tension of the nonwetting liquid, water, and θ is the contact angle of water. The pores generally are not of spherical shape or a constant diameter. They usually vary in their form and size. Thus, statements of any pore diameter are always to be viewed with the above in mind.

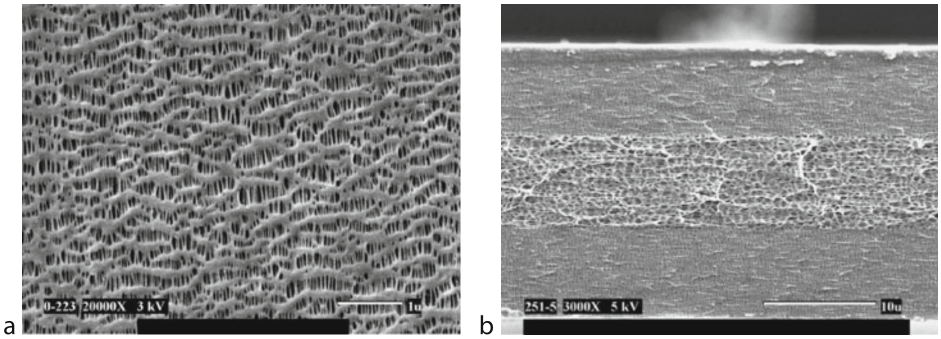
Another technique, Capillary Flow Porometry has been developed by Porous Materials Inc. [98] to characterize battery separators [99, 100]. The instrument can measure a number of characteristics of battery separators such as size of the pore at its most constricted part, the largest pore size, pore-size distribution, permeability, and envelope surface area [101].

Scanning electron microscopy (SEM) is also used to examine separator morphology. SEM pictures of some commercial membranes are shown in Figs. 4–6.



Rechargeable Batteries, Separators for. Figure 4

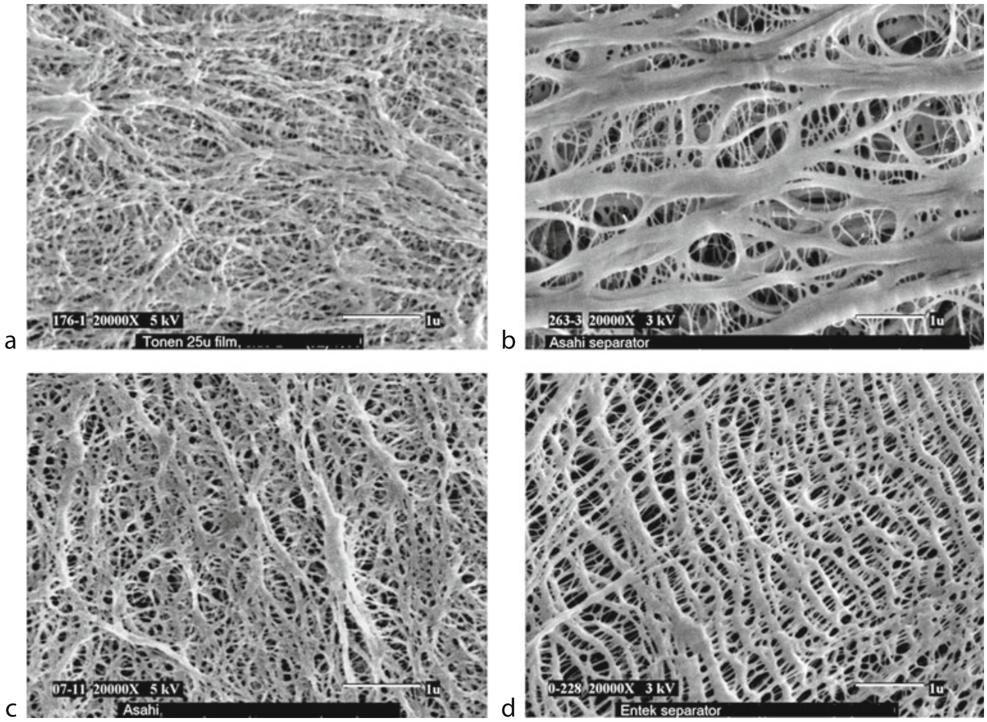
Scanning electron micrographs of surface of single-layer Celgard separators used in lithium batteries (**a**) 2400 (PP), (**b**) 2500 (PP), and (**c**) 2730 (PE)



Rechargeable Batteries, Separators for. Figure 5

Scanning electron micrographs of Celgard trilayer separator (2325) used in lithium-ion batteries (**a**) surface SEM; (**b**) cross-section SEM

The surface SEM of Celgard 2400, 2500, and 2730 are shown in Fig. 4. It is clear from the images that the pores are uniformly distributed. Both Celgard 2400 and 2500 are single-layer PP separators, but the pore size of Celgard 2500 is substantially larger than Celgard 2400. Thus, it has lower resistance and is more suited for high rate applications. Fig. 5 shows the surface SEM and cross-section SEM of Celgard 2325. The surface SEM only shows the PP pores while the PE pores are visible in the cross-section. It is clear from the image that all three layers are of equal thickness. The SEM pictures of separators made by wet process are shown in Fig. 6.



Rechargeable Batteries, Separators for. Figure 6
Scanning electron micrographs of separators made by wet process and used in lithium-ion batteries (**a**) Setela (Tonen), (**b**) Hipore–1 (Asahi), (**c**) Hipore-2 (Asahi), and (**d**) Teklon (Entek)

The pore structure of all of these membranes is very similar. Asahi-1 (Fig. 6b) separator has significantly larger pores compared to the other membranes.

Image analysis has been used to characterize the pore structure of synthetic membrane materials [102]. The films can also been characterized by scanning tunneling microscopy, atomic force microscopy, and field emission scanning electron microscopy [49, 103]. The pore size of the membranes can also be calculated from Eq. 5, once the MacMullin number and Gurley values are known.

Puncture Strength A separator is required to have sufficient physical strength to endure the rigors of cell assembly and day-to-day charge–discharge cycling. Physical strength is required to withstand basic handling, cell blocking/assembly, physical shock, punctures, abrasion, and compression.

The Puncture Strength (PS) is the weight that must be applied to a needle to force it completely through a separator [55, 104]. It has been used to indicate the tendency of separators to allow short circuits in a cell that may occur due to holes generated in the separator by the rough surface of an electrode during the battery assembly and charge–discharge cycle. The PS requirement for lithium-ion batteries is higher them lithium-foil batteries, because the separator must contend with two rough surfaces. Commercially available puncture strength machines made for textiles tend to give meaningless results when testing battery separator membranes. More reproducible results can be obtained with a load frame (such as an Instron Machine). The mix penetration strength is a better measure of mechanical strength for battery separators as it measures the force required to create a short through the separator when electrode mix is pushed through it.

The strength of the separator depends greatly on the materials used and the manufacturing method. The wet-biaxial method simultaneously stretches in the

MD and TD directions and thus achieves a material that has tensile modulus and rupture strength in both directions. Both high polymer entanglement and stretching help increase the physical strength of the separator.

Mix Penetration Strength The force required to create a short through a separator due to mix (electrode material) penetration defines mix penetration strength. In this test, force (with a ½-inch diameter ball) is applied on the positive electrode/separator/negative electrode sandwich and the force at which the mix penetrates through the separator and creates an electronic short is called mix penetration force. Mix penetration strength is used to indicate the tendency of separators to allow short circuits during battery assembly. The mix penetration resistance test is more closely related to particle penetration resistance compared to puncture resistance [45].

Tensile Strength The tensile strength measurements (e.g., Young's Modulus, Percent offset strength, elongation at break, stress at break) can be made by utilizing widely known standard procedures. These tests are carried out in both MD and TD directions. The tensile properties are dependent on the manufacturing process. The Uniaxially oriented films have high strength in only one direction, whereas biaxially oriented films are more uniformly strong in both MD and TD directions. ASTM test method D882-00 "Standard test method for tensile properties of thin plastic sheeting" is an appropriate test.

The separator should be strong enough to withstand mechanical handling during cell winding and assembly. It should be dimensionally stable and should not neck down during winding. The decrease in width will allow the electrodes to touch each other and create a short. Thus, the tensile property of the separator should be very strong in MD direction compared to TD direction.

Shrinkage Shrinkage test is carried out on both MD and TD directions. In this test, the dimensions of separators are measured and then stored at 90°C for a fixed time. The shrinkage is then calculated from the change in dimensions as shown in Eq. 10,

$$\text{Shrinkage}(\%) = \frac{L_i - L_f}{L_i} \times 100 \tag{10}$$

where L_i is the initial length and L_f is the final length of separator after high temperature storage. The uniaxially stretched separators tend to shrink in MD direction only, while the biaxially stretched separators shrink in both MD and TD directions. The shrinkage of separators can also be compared by carrying out Thermal Mechanical Analysis (TMA) test at a constant load and rate.

Shutdown Separator shutdown is a useful and essential mechanism for limiting temperature and preventing venting in short-circuited cells [42]. It usually takes place close to the melting temperature of the polymer when the pores collapse turning the porous ionically conductive polymer film into a nonporous insulating layer between the electrodes. At this temperature, a significant increase in cell impedance occurs and passage of current through the cell is restricted. This prevents further electrochemical activity in the cell, thereby shutting the cell down before an explosion can occur.

The ability of the PE-based separator to shutdown the battery is determined by its molecular weight, percent crystallinity (density), and process history. Material properties and processing methods might need to be tailored so that the shutdown response is spontaneous and complete. The optimization needs to be done without affecting the mechanical properties of the material in the temperature range of interest. This process is easier to do with the trilayer separators, since one material is utilized for the shutdown response and another for the mechanical properties. Polyethylene containing separators, in particular trilayer laminates of polypropylene, polyethylene, and polypropylene, appear to have the most attractive properties for preventing thermal runaway in lithium-ion cells [101–105]. The shutdown temperature of 130°C is usually sufficient to control the cell heating and avoid thermal runaway in lithium-ion cells. A lower temperature shutdown will be desirable if it does not affect the separator mechanical properties or high temperature cell performance in any adverse way.

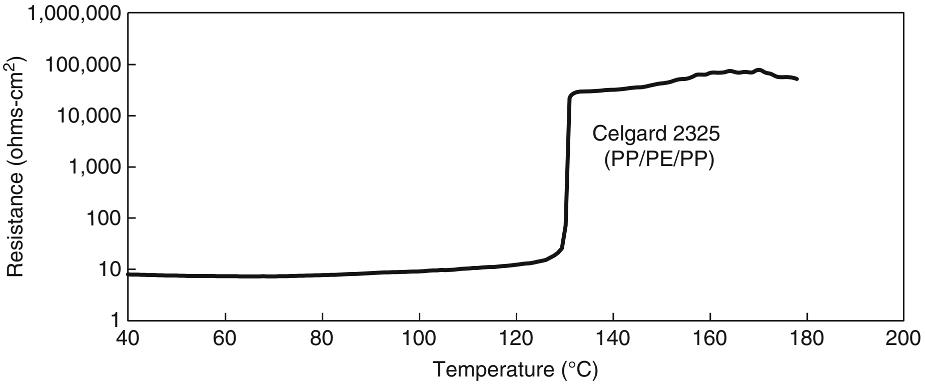
The shutdown property of separators is determined by measuring the impedance of a separator while the temperature is linearly increased [42, 89]. Figure 7 shows actual measurement for Celgard® 2325 membrane. The heating rate was around 60°C/min and the impedance was measured at 1 kHz. The rise in impedance corresponds to a collapse in pore structure due to melting of the separator. A 1,000-fold increase in impedance is necessary for the separator to stop thermal runaway in the battery. The drop in impedance corresponds to opening of the separator due to coalescence of the polymer, and/or to penetration of the separator by the electrodes; this phenomenon is referred to as a loss in "melt integrity." This test is fairly reliable in indicating the temperature at which the impedance rises, but some variability in characterizing the subsequent drop in impedance may occur.

In Fig. 7, the shutdown behavior of a multilayer (PP/PE/PP) separator (Celgard 2325) is shown. The impedance rise occurred near the melting point of polyethylene (130°C) and remained high until such time as the melting point of polypropylene (165°C) is attained. The shutdown temperature of the separator is governed by the melting point of the separator material. At the melting point, the pores in the separator collapse to the form a relatively nonporous film between the anode and the cathode. This was confirmed by DSC. The DSC scan in Fig. 8 gives a peak melting temperature of 135°C for Celgard 2730, 168°C for Celgard 2400, and 135/165°C for Celgard 2325. The shutdown behavior of thinner separators (<20 µm) is very similar to thicker separators. The battery manufacturers have been very successful in using the thinner separators without compromising on the shutdown behavior of the separators.

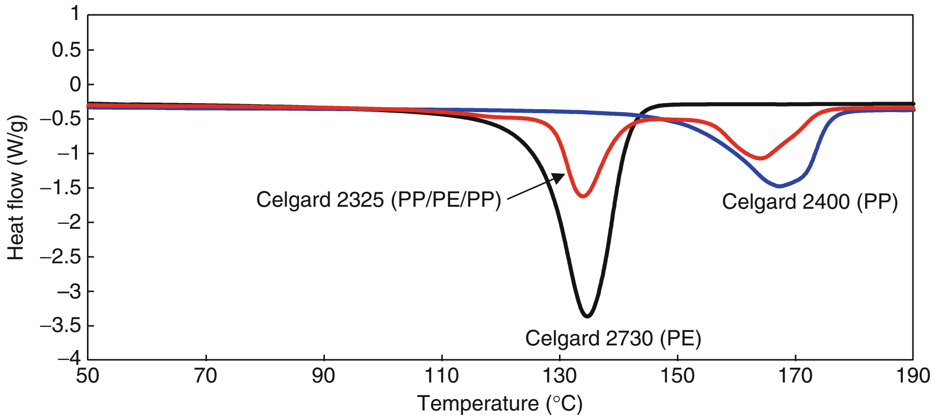
Laman et al. introduced the use of impedance measurements as a function of temperature to characterize shutdown separators [89]. Using a temperature scan rate of 1°C/min they found that the impedance increased several orders of magnitude near the melting point of the separator. They verified the patent claims of Lundquist et al. [106] that bilayer separators of PE and PP gave a temperature window of high impedance extending approximately between the melting point of the polymers. The concept of using separators consisting of distinct layers, one of which could act as a fuse, was developed by Lundquist et al. [106, 107]. Laman's results have been corroborated by Geiger et al. [38] and Spotnitz et al. [94]. Spotnitz et al. developed a thin layer cell which allowed temperature scan rates of 5°C/min and higher and obtained results similar to those of Laman et al.

Prior work related with shutdown separators also involved application of waxes on membranes [108, 109]. In these cases, the wax or low melting polymers were coated on the polyolefin separator. The disadvantage of this technique is that the coating can block the pores of the separator and thus can affect the performance by increasing separator resistance. Moreover, the coating level has to be very high to get complete shutdown.

The shutdown characteristic provides protection from external short circuit and during cell overcharge. It provides little protection from internal shorts should



Rechargeable Batteries, Separators for. Figure 7
Internal impedance (at 1 kHz) of Celgard 2325 (PP/PE/PP) separator as a function of temperature. Heating rate: 60°C/min



Rechargeable Batteries, Separators for. Figure 8
DSC of Celgard 2730 (PE), 2400 (PP), and 2325 (PP/PE/PP)

they occur. Should the electrodes touch each other or become shorted from a dendritic growth of soluble impurity or other dendrite forming soluble material, the separator only helps in avoiding delayed failures. In case of an instant failure during internal short circuit, the heating rate is too high and the separator shutdown is not fast enough to control the heating rate.

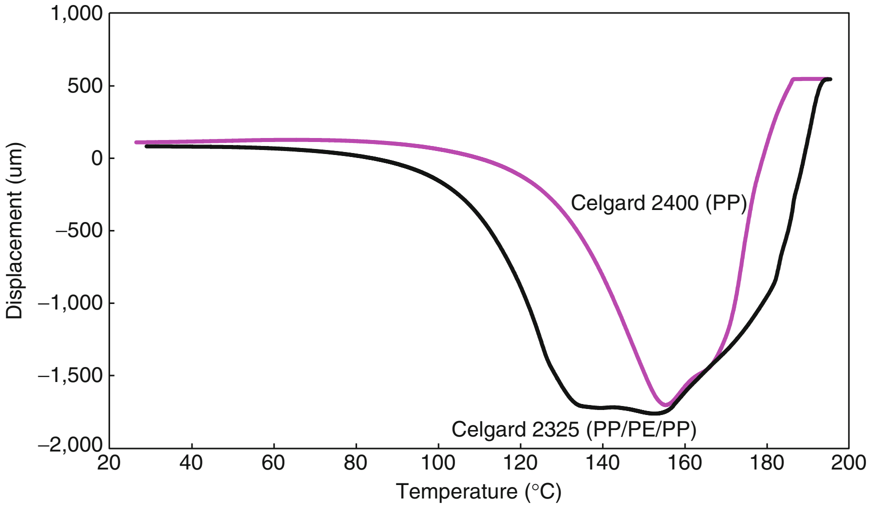
Melt Integrity The separators used in lithium-ion batteries should have high temperature melt integrity. The separator should maintain its melt integrity after shutdown so that the electrodes do not touch and create a short. This helps in avoiding the thermal runaway even when the cell is exposed to high temperatures. Thermal Mechanical Analysis (TMA) is a very good technique to measure the high temperature melt integrity of separators.

TMA involves measuring the shape change of a separator under load while the temperature is linearly increased. Typically, separators show some shrinkage, and then start to elongate and finally break as shown in Fig. 9. This test utilizes a small separator sample (about 5–10 mm length (MD) and about 5 mm width), which is held in mini-instron-type grips. The sample is held with a constant 2 g load while the temperature is ramped at 5°C/min past the melting point until the tension ruptures the film. Three parameters are reported from TMA test – shrinkage onset temperature, melt temperature, and melt rupture temperature. TMA has proven to be a more reproducible measure of melt integrity of the separator [107].

Figure 9 shows the TMA data for two different Celgard membranes. The shrinkage onset temperature, deformation temperature, and rupture temperature are summarized in Table 6. The single-layer PP membrane (Celgard 2400) showed a higher softening temperature (~121°C), a deformation temperature around 160°C, and a very high rupture temperature around 180°C. The multilayer polypropylene/polyethylene/polypropylene separator (Celgard 2325) combined the low-temperature shutdown property of polyethylene with the high temperature melt integrity of polypropylene, resulting in a separator with softening (~105°C) and melt temperature (~135°C) very similar to PE and rupture temperature (~190°C) very similar to PP.

Separators with melt integrity greater than 150°C are desirable for lithium-ion cells. The Trilayer separators with polypropylene on the outside helps in maintaining the melt integrity of the separators at higher temperatures compared to single-layer PE separators. The choice of a shutdown separator for bigger lithium-ion cells being developed for hybrid and electric vehicles is highly specific to the design of the cell.

Wettability and Wetting Speed Two physical properties of separators, which are important to the operating characteristics of a battery, are electrolyte absorption and electrolyte retention. Any good separator should be able to absorb a significant amount of electrolyte and also retain the absorbed electrolyte when the cell is in operation. These are more important in sealed cells where no free electrolyte is present.



Rechargeable Batteries, Separators for. Figure 9
TMA of Celgard 2400 (PP) and 2325 (PP/PE/PP). A constant load (2 g) is applied while the temperature is ramped at 5°C/min

Rechargeable Batteries, Separators for. Table 6 TMA data for typical Celgard separators

	Celgard 2730	Celgard 2400	Celgard 2325
Shrinkage onset temperature (°C)	100	121	106
Deformation temperature (°C)	125	156	135, 154
Rupture temperature (°C)	140	183	192

A maximum amount of electrolyte in the separator is desirable to achieve minimum cell internal resistance. Quite often separator dry-out or lack of sufficient amount of electrolyte is misinterpreted as inadequate wetting of the separator.

The separator wettability can limit the performance of batteries by increasing the separator and cell resistance. Separator wetting speed can be correlated with electrolyte filling time in real cells. The wetting speed is determined by the type of polymer (surface energy), pore size, porosity, and tortuosity of the separators. There is no generally accepted test for separator wettability. However, a simple wicking test by placing a drop of electrolyte onto the separator is a good indication of wettability. Standard dyne-solutions of known surface tension values exist and can be used to adjust the surface tension of the electrolyte as to expedite wetting. The contact angle is also a good measure of wettability. The uptake of electrolyte by many hydrophobic polymer separators can be enhanced either by wetting agents or ionic-functional groups (e.g., ion-exchange membranes).

Effect of Separator on Cell Performance and Safety
Although the material of a battery separator is inert and does not influence electrical energy storage or output, its physical properties greatly influence performance and safety of the battery. This is especially true for lithium-ion cells and, thus, the battery manufacturers have started paying more attention to separators while designing the cells. The cells are designed in such a way that separators do not limit the performance, but if the separator properties are not uniform, or if there are other issues, it can affect the performance and safety of cells. This section will focus on the effect of the separator properties on cell performance and safety. Table 7 shows different types of safety and performance tests for lithium-ion batteries and the corresponding important separator property and how it affects performance and/or safety.

To achieve good performance of lithium-ion cells, the separators should have low resistance and low shrinkage and strength across the thickness.

Rechargeable Batteries, Separators for. Table 7 Safety and performance tests for lithium-ion batteries and the corresponding important separator property and its affect on the cell performance and/or safety

Cell property	Separator property	Comments
Cell capacity	Thickness	Cell capacity can be increased by making the separator thinner
Cell internal resistance	Resistance	Separator resistance is a function of thickness, pore size, porosity, and tortuosity
High rate performance	Resistance	Separator resistance is a function of thickness, pore size, porosity, and tortuosity
Fast charging	Resistance	Low separator resistance will aid in overall faster charging by allowing higher and/or longer constant current charging
High temperature storage	Oxidation resistance	Oxidation of separators can lead to poor storage performance and reduce performance life
High temperature cycling	Oxidation resistance	Oxidation of separators can lead to poor cycling performance
Self-discharge	Weak areas, pinholes	Soft shorts during cell formation and testing can lead to internal current leakage
Long-term cycling	Resistance, shrinkage, pore size	High resistance, high shrinkage, and very small pore size can lead to poor cycling performance
Overcharge	Shutdown behavior; high temperature melt integrity	Separator should completely shutdown and then maintain its melt integrity at high temperatures
External short circuit	Shutdown behavior	Separator shutdown stops the cells from overheating
Hotbox	High temperature melt integrity	Separator should be able to keep the two electrodes apart at high temperatures
Nail crush	Shutdown (to stop delayed failure)	In case of internal shorts, the separator may be the only safety device to stop the cell from overheating.
Bar crush	Shutdown (to stop delayed failure)	In case of internal shorts, the separator may be the only safety device to stops the cell from overheating.

The separator with high resistance will perform poorly during high rate discharge and will also increase the cell charging time. Low shrinkage is a very important characteristic for separators, especially for higher capacity cells. These cells are used in high-speed laptop computers, which can experience higher temperatures (~70–75°C) under certain conditions [110]. This can lead to shrinkage of separators and ultimately higher cell resistance and poor long-term cycling. The shrinkage in TD direction can lead to safety issues because of an internal short between the electrodes. Larger pores can lead to shorts during cell manufacturing or can fail during Hipot testing. Larger pores will allow more soft shorts and higher self-discharge, especially during high temperature storage. Very small pore size can lead to higher resistance and poor cycle life during high temperature cycling and storage. Thus, the pore size of the separator should be optimized to achieve good strength and performance.

One of the ways to increase cell capacity is by decreasing the thickness of separators. The newer high capacity cells (>2.6 Ah) generally use 16 and 12 μm separators as compared to 20–25 μm separators used in cells with 2.2–2.4 Ah capacity. The thinner separators offer lower resistance and help in increasing the capacity; but the amount of electrolyte they can hold is less and their mechanical strength is often not as high as thicker separators. Thus, appropriate changes should be made in cell design to keep the cell safe. The handling and manufacturing of thinner separators is also a challenge for the separator manufacturers. They are required to maintain the same electrical and mechanical properties and better quality for thinner separators. The separator manufacturers have installed better controls and quality standards. Many battery experts are of the opinion that the 16 µm is the thinnest that can be used while still maintaining the stringent performance and safety requirements of lithium-ion cells. The use of thinner separators has often resulted in voluminous battery recalls [111].

The separators inside the lithium-ion batteries experience extreme oxidizing environment on the side facing the positive electrode and extreme reducing environment on the side facing the negative electrode. The separators should be stable in these conditions during long-term cycling especially at high temperatures. Separators with poor oxidation resistance can lead to poor high temperature storage performance and poor long-term cycling behavior. The oxidation resistance properties of trilayer (PP/PE/PP) separators with PP as the outside layer and PE as inner layer is superior compared to polyethylene separators. This is because of the better oxidation resistance properties of polypropylene in contact with the positive electrode in a lithium-ion cell.

The products formed by the decomposition of the electrolyte can also block the pores of the separator, leading to increase in cell resistance. The separators with lower resistance also help in better low-temperature performance. At very low temperatures, the resistance of the electrolytes is very high and thus smaller contribution from separator helps in keeping the cell resistance low.

The lithium-ion cells have demonstrated power loss when aged and/or cycled at high temperatures. Norin et al. [112] demonstrated that the separator is at least partly responsible for the power loss due to the intrinsic increase in its ionic resistance. They showed that impedance increased significantly upon cycling and/or aging of lithium-ion cells at elevated temperatures and that separators account for about 15% of the total cell impedance rise. They later reported that the loss in ionic conductivity of the separator was due to blocking of the separator pores with the products formed due to electrolyte decomposition, which was significantly accelerated at elevated temperatures [113].

There are several groups that regulate, or provide testing, to verify safe operation of lithium-ion cells under abuse conditions. The US Department of Transportation (DOT) classifies all lithium-ion batteries as hazardous materials for shipping in the same category as lithium metal primary batteries [114]. The DOT grants exceptions based on the cell capacity and ability of the cells to pass specified tests. In addition, the UL Laboratories [115, 116], the International Electrotechnic Commission [117], and the United Nations (UN) [118] have developed standardized safety testing procedures. These tests are designed to assure that cells are safe to ship and are resistant to typical abuse conditions such as internal shorting, overcharge, overdischarge, vibration, shock, and temperature variations that may be encountered in normal transportation environments.

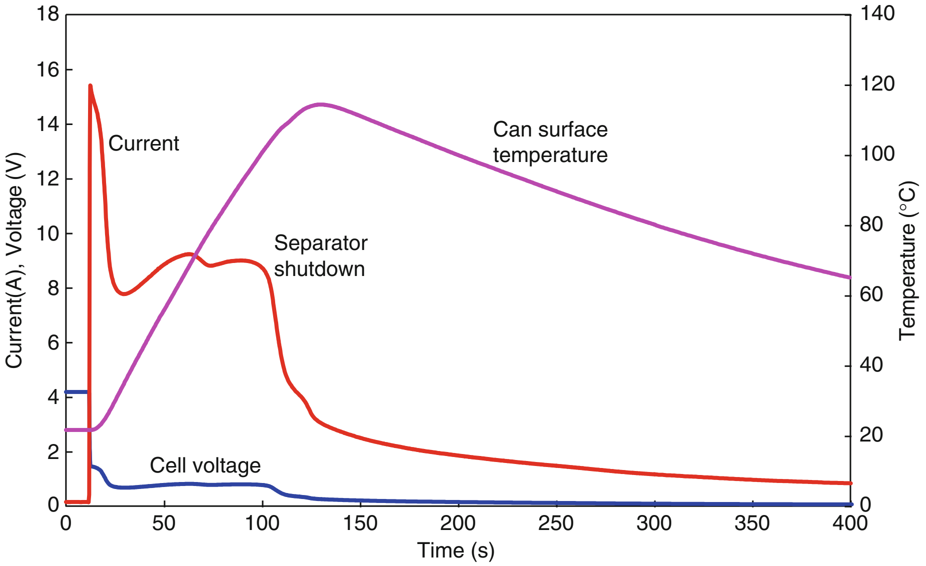
Underwriters Laboratories (UL) requires that consumer batteries pass a number of safety tests (UL 1642 [119] and UL-2054 [120]). There are similar recommendations from UN for transport of dangerous goods, [121] the International Electrotechnical Commission (IEC), and the Japan Battery Association [122]. An abnormal increase in cell temperature can occur from internal heating caused by either electrical abuse – overcharge or short circuit – or mechanical abuse – nail penetration or crush. Higher cell temperature could also be a result of external heating. For this reason, lithium-ion cells used in battery packs are designed with safety control circuits that have redundant safety features (PTC, CID, vent, thermal fuse, etc.). Shutdown separators are one of the safety devices inside the cell and act as a last line of defense. The separator shutdown is irreversible, which is fine for polyethylene-based separators, which melt around 130°C.

The impedance of the separator increases by two to three orders of magnitude due to an increase in cell temperature, resulting from cell abuse (e.g., short circuit, overcharge). The separator should not only shutdown around 130°C, but it should also maintain its mechanical integrity at higher temperatures, preferably at temperatures as high as 200°C. If the separator does not shutdown properly then the cell will continue to heat during an overcharge test and can lead to thermal runaway. The high temperature melt integrity of separators is also a very important property to keep the cell safe during extended overcharge or during extended exposure to higher temperatures.

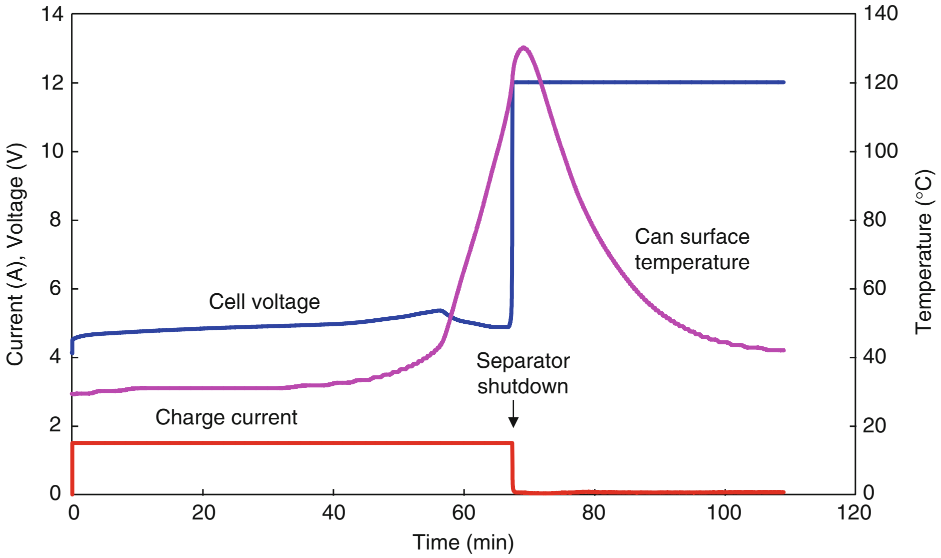
Figure 10 shows a typical short-circuit curve for an 18650 lithium-ion cell with shutdown separator, $LiCoO_2$ positive electrode, and MCMB carbon negative electrode. For the results shown in Fig. 10, the cell tested did not have other safety devices (e.g., CID, PTC), which usually work before separator shutdown. As soon as the cell is short circuited externally through a shunt resistor, the cell starts heating because of the large current drained through the cell. The shutdown of the separator, which occurs around 130°C, stops the cell from heating further. The current decrease is caused by increase of battery internal resistance due to separator shutdown. The separator shutdown helps in avoiding the thermal runaway of the cell.

Cells can be overcharged when the cell voltage is incorrectly detected by the charging control system, or when the charger breaks down. When this happens, the lithium ions remaining in the cathode are removed and more lithium ions are inserted into the anode then under standard charging conditions. If the lithium insertion ability of the carbon anode is limited, lithium metal in the form of dendrites may be deposited on the carbon and cause a drastic reduction in thermal stability. At higher charging rates, the heat output increases greatly because the joule heat output is proportional to I^2R. Several exothermic reactions (e.g., reaction between lithium and electrolyte, thermal decomposition of anode and cathode, thermal decomposition of electrolyte, etc.) occur inside the cell as its temperature increases. Separator shutdown happens when cell temperature reaches melting point of polyethylene as shown in Fig. 11. The CID and PTC of the 18650 cells were removed, to identify the role of the separator. The current decrease is caused by increase of battery internal resistance due to separator shutdown. Once the pores of the separator have closed due to softening, the battery cannot continue to be charged or discharged, and thus thermal runaway is prevented. During continued overcharge, the separator should maintain its shutdown feature and should not allow the cell to heat again. It should also maintain its melt integrity and should not allow the two electrodes to touch each other.

The separator should also not allow any dendrite to penetrate through the separator to avoid internal



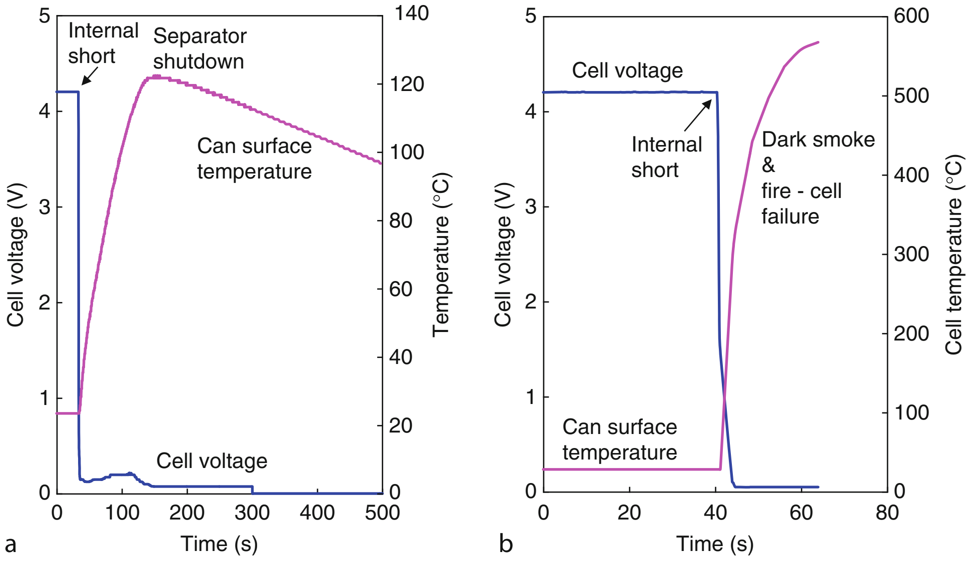
Rechargeable Batteries, Separators for. Figure 10
Typical short-circuit behavior of a 18650 lithium-ion cell with shutdown separator and without PTC (positive temperature coefficient) and CID (current interrupt device). This test simulates external short circuit of cell



Rechargeable Batteries, Separators for. Figure 11
Typical overcharge behavior of a 18650 lithium-ion cell with shutdown separator. The PTC (Positive Temperature Coefficient) and CID (Current Interrupt Device) were removed from the cell header

shorts. During an internal short, separator is the only safety device, which can stop the thermal runaway. If the heating rate is not too high then the separator shutdown can help in controlling the heating rate and stop thermal runaway. In a nail penetration test, an instantaneous internal short results the moment the nail penetrates into battery. Enormous heat is produced from current flow (double-layer discharge and electrochemical reactions) in the circuit by the metal nail and electrodes. Contact area varies according to depth of penetration. In general, the shallower the penetration depth, the smaller the contact area and therefore the greater are the local current density and heat production. Thermal runaway is likely to take place as local heat generation induces electrolyte and electrode materials to decompose. On the other hand, if the battery is fully penetrated, the increased contact area would lower the current density, and the cell could pass the nail penetration test. A detailed investigation of various internal short-circuit scenarios is presented by Santhanagopalan et al. [229]. Internal short-circuit tests are more difficult to pass than the external short-circuit tests described earlier, because the nature of the short cannot be determined a priori.

Figure 12 shows the typical nail penetration behavior of an 18650 lithium-ion cell with shutdown separator, $LiCoO_2$ positive electrode, and MCMB carbon negative electrode. Clearly, there was a voltage drop from 4.2 to 0.0 V, instantaneously, as the nail penetrated through (when internal short circuit occur), and temperature rose. When the heating rate is low the cell stops heating when the temperature is close to separator shutdown temperature as shown in Fig. 12a. If the heating rate is very high, then the cell continues to heat and fails the nail penetration test as shown in Fig. 12b. In this case, the separator shutdown is not fast enough to prevent thermal runaway. Thus, separator only helps in avoiding delayed failures in case of internal short circuit as simulated by nail and bar crush tests. Separators with high temperature melt integrity and good shutdown feature (to avoid delayed failures) are needed to pass internal short-circuit test. Thinner separators (<20 μm) used in high capacity cells should offer similar shutdown and high temperature melt integrity properties as thicker separators. The decrease in separator strength should be balanced with changes in cell design. The separator properties across the length and width should be very uniform to keep the cell safe during abnormal use.



Rechargeable Batteries, Separators for. Figure 12
Typical nail penetration behavior of an 18650 lithium-ion cell with shutdown separator. This test simulates internal short circuit of a cell. (**a**) Cell passed nail penetration test; (**b**) cell failed nail penetration test

The mechanism and characteristics of thermal cut-off devices in several prismatic lithium-ion cells was studied by Venugopal [124] by monitoring the impedance at 1 kHz and the Open Circuit Voltage (OCV) of the cells as a function of temperature. All the cells studied contained PE-based separators with a shutdown temperature between 130°C and 135°C. Within this narrow temperature range, the shutdown separators caused a sharp and irreversible rise in impedance of the cell. Single-layer PE separators were effective up to around 145°C, above which they demonstrated a meltdown effect. Trilayer separators had meltdown temperatures as high as 160°C because of the presence of additional layers of higher melting PP. It was found that the separators, alone, are not able to shutdown the cell completely. In case of an overcharged test, the cell could continue to charge at lower currents even after the shutdown event, rendering the cell a potential hazard if not disposed of immediately and safely. This usually does not become an issue in commercial cells because the cell manufacturers have addressed this issue by including multiple cut-off devices within a single cell. The use of inorganic coatings and fillers has also drawn considerable attention in order to prevent cell failure in such cases [78, 125].

Development efforts are under way to displace the use of microporous membranes as battery separators and instead use gel or polymer electrolytes. Polymer electrolytes, in particular, promise enhanced safety by eliminating organic volatile solvents. The next two sections are devoted to solid polymer and gel-polymer-type lithium-ion cells with focus on their separator/electrolyte requirements.

Separator for Lithium Polymer Batteries

Due to their high theoretical capacity, lithium polymer batteries have long been identified as a very promising technology to meet the requirements of upcoming applications such as standby power and electric vehicles. Research and development of polymer electrolytes for ambient-temperature rechargeable lithium batteries has always been very active. Rapid progress for the past two decades in this field has led to numerous monographs and reviews [126–131]. These polymers are generally polyethers, poly(ethylene oxide) (PEO), or poly(propylene oxide) (PPO).

Solid polymer electrolytes serve two principal roles in rechargeable lithium batteries. Not only do they function as the traditional electrolyte, i.e., the medium for ionic transport, but also as the separator which insulates the cathode from the anode. Consequently, the polymer electrolyte must have sufficient mechanical integrity to withstand electrode stack pressure and stresses caused by dimensional changes, which the rechargeable electrodes undergo during charge/discharge cycling.

Lithium polymer electrolytes formed by dissolving a lithium salt LiX (where X is preferably a large soft anion) in poly(ethylene oxide) PEO can find useful application as separators in lithium rechargeable polymer batteries [132–134]. Thin films must be used due to the relatively high ionic resistivity of these polymers. For example, the lithium-ion conductivity of PEO-Li salt complexes at 100°C is still only about 1/100 the conductivity of a typical aqueous solution.

A polymer electrolyte with acceptable conductivity, mechanical properties, and electrochemical stability has yet to be developed and commercialized on a large scale. The main issues that must be resolved for a completely successful operation of these materials are the reactivity of their interface with the lithium metal electrode and the decay of their conductivity at temperatures below 70°C. Croce et al. found an effective approach for reaching both of these goals by dispersing low-particle-size ceramic powders in the polymer electrolyte bulk [135, 136]. They claimed that this new "nanocomposite polymer electrolytes" had a very stable lithium electrode interface and an enhanced ionic conductivity at low temperature, combined with good mechanical properties. Fan et al. [137] has also developed a new type of composite electrolyte by dispersing fumed silica into low to moderate molecular weight PEO.

The gel-type polymer electrolyte prepared by dispersing ceramic powders (e.g., Al_2O_3) into a matrix formed by a lithium salt solution contained in a poly(acrylonitrile) (PAN) network was reported by Appetecchi et al. [138] These new types of composite gel electrolytes had high ionic conductivity, wide electrochemical stability, and particularly, high chemical integrity even at temperatures above ambient. Kim et al. [139] used a blend of PVdF-HFP and PAN as a matrix polymer to attain high ionic conductivity and good mechanical strength. The PAN can give mechanical integrity and structural rigidity to a porous membrane without inorganic fillers. The high ionic conductivity was due to the high volume of pores and a high affinity of the membrane for electrolyte solution [140].

Separator for Lithium-Ion Gel Polymer Batteries

The solid polymer electrolyte approach provides enhanced safety, but the poor ambient-temperature conductivity excludes their use for battery applications, which requires good ambient-temperature performance. In contrast, the liquid lithium-ion technology provides better performance over a wider temperature range, but electrolyte leakage remains a constant risk. Midway between the solid polymer electrolyte and the liquid electrolyte is the "hybrid polymer" electrolyte concept leading to the so-called gel polymer lithium-ion batteries. Gel electrolyte is a two-component system, namely, a polymer matrix swollen with a liquid electrolyte. The gel polymer electrolyte approach to the lithium-ion technology combines the positive attributes of both the liquid (high ionic conductivity) and solid polymer electrolytes (elimination of leakage problems).

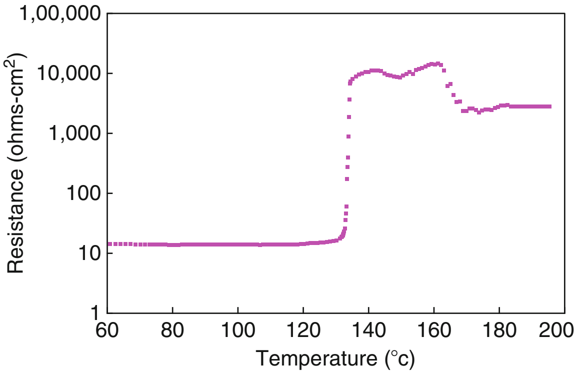
Gel polymer lithium-ion batteries replace the conventional liquid electrolytes with an advanced polymer electrolyte membrane. These cells can be packed in light-weight plastic packages as they do not have any free electrolyte and they can be fabricated in any desired shape and size. They are now increasingly becoming an alternative to liquid electrolyte lithium-ion batteries, and several battery manufacturers, such as Sanyo, Sony, and Panasonic, have started commercial production [141, 142]. Song et al. [143] have recently reviewed the present state of gel-type polymer electrolyte technology for lithium-ion batteries. They focused on four plasticized systems, which have received particular attention from a practical viewpoint, i.e., poly (ethylene oxide) (PEO), poly (acrylonitrile) (PAN) [144], poly (methyl methacrylate) (PMMA) [145, 146], and poly (vinylidene fluoride) (PVdF) based electrolytes [147–150].

One particular version of the lithium-ion gel polymer cells, also known as plastic lithium-ion cell (PLION™), was developed by Bellcore [151–153]. In this case, Gozdz et al. developed a microporous plasticized PVdF-HFP-based polymer electrolyte that served both as separator and electrolyte. In PLION™ cells, the anode and cathode are laminated onto either side of the gellable membrane. Good adhesion between the electrodes and the membranes is possible because all three sheets contain significant amounts of a PVdF copolymer that can be melted and bonded during the lamination step.

The PVdF-HFP separators used in PLION™ cells were around 3 ml thick, and had poor mechanical properties. It has been reported that the major source of rate limitation in PLION™ cells was the separator thickness [154]. The rate capability of these cells can be significantly improved by decreasing the separator thickness to that typically used in liquid electrolyte system. Moreover, in the absence of shutdown function, the separator does not contribute to cell safety in any way. Park et al. reported that the HFP content in separators did not have any significant impact on cell performance [155]. The Bellcore process has proven to be an elegant laboratory process but is difficult to implement in large-scale production.

To overcome the poor mechanical properties of polymer- and gel-polymer-type electrolytes, microporous membranes impregnated with gel polymer electrolytes, such as PVdF, PVdF-HFP, and other gelling agents, have been developed as an electrolyte material for lithium batteries [156–166]. Gel-coated and/or gel-filled separators have some characteristics that may be harder to achieve in the separator-free gel electrolytes. For example, they can offer much better protection against internal shorts when compared to gel electrolytes and can therefore help in reducing the overall thickness of the electrolyte layer. In addition the ability of some separators to shutdown at a particular temperature allows safe deactivation of the cell under overcharge conditions.

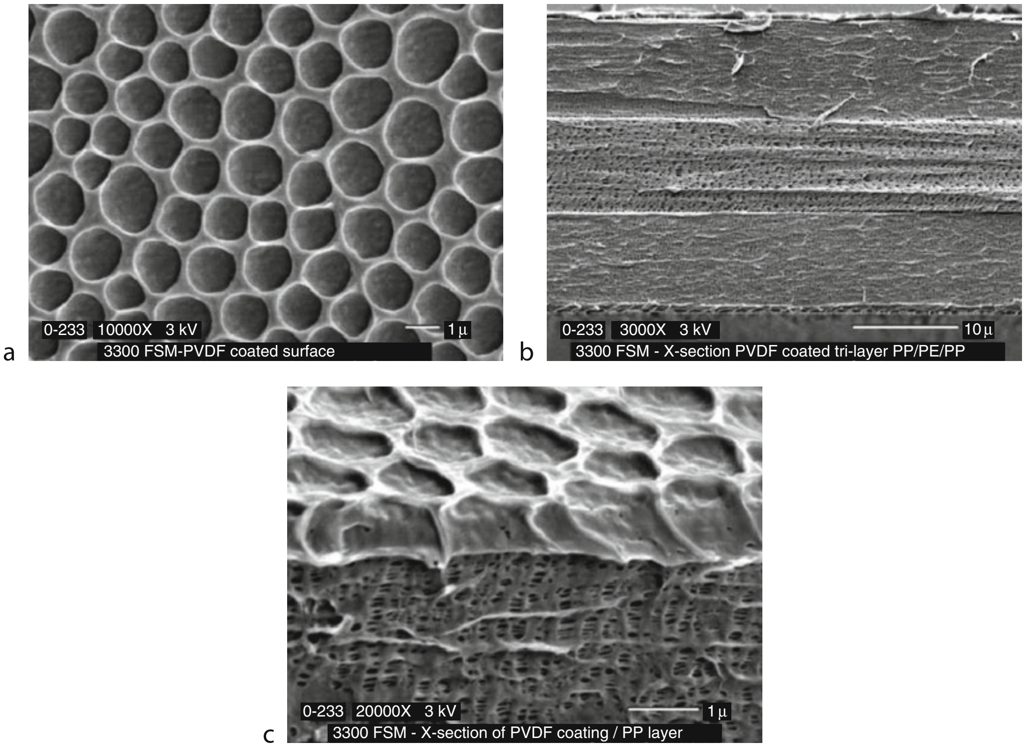
The shutdown behavior of PVdF-coated Celgard a trilayer membrane is shown in Fig. 13. The shutdown is defined by the sharp increase in resistance around 130°C. The PVdF coating should be porous and should not block the pores to maintain similar ionic conductivity. The scanning electron micrographs of PVdF-coated membrane is shown in Fig. 14. The cross-section SEM of Celgard 3300 provides visual evidence that the coating is porous and is not blocking the pores of the top PP layer.



Rechargeable Batteries, Separators for. Figure 13 Internal impedance (at 1 kHz) of PVdF-coated Celgard trilayer separators as a function of temperature. Heating rate: 60°C/min

Abraham et al. [158] were the first ones to propose saturating commercially available microporous polyolefin separators (e.g., Celgard®) with a solution of lithium salt in a photopolymerizable monomer and a nonvolatile electrolyte solvent. The resulting batteries exhibited low discharge rate capability due to the significant occlusion of the pores with the polymer binder and the low ionic conductivity of this plasticized electrolyte system. Dasgupta and Jacobs [157, 168] patented several variants of the process for the fabrication of bonded-electrode lithium-ion batteries, in which a microporous separator and electrode were coated with a liquid electrolyte solution, such as ethylene-propylene-diene (EPDM) copolymer and then bonded under elevated temperature and pressure conditions. This method required that the whole cell assembling process be carried out in scrupulously anhydrous conditions, which make this approach difficult, and expensive.

The later methods, proposed by Motorola [159, 170] and Mitsubishi Electric [171] researchers, differ


a
0-233 10000X 3 kV
1 µ
3300 FSM-PVDF coated surface
b
0-233 3000X 3 kV
10 µ
3300 FSM - X-section PVDF coated tri-layer PP/PE/PP
c
0-233 20000X 3 kV
1 µ
3300 FSM - X-section of PVDF coating / PP layer

Rechargeable Batteries, Separators for. Figure 14
Scanning Electron Micrographs of Celgard PVdF-coated separators used in lithium gel polymer batteries (**a**) surface SEM, (**b**) cross-section SEM of coated layer, and (**c**) cross-section of PVdF coating

in implementation details, but they share a common feature in that a separate adhesive layer (PVdF) is applied to the separator and used to bond the electrode and the separator films, using in the first case the hot, liquid electrolyte as an in situ PVdF plasticizer. Sony [172, 173] researchers described the use of thin, liquid electrolyte-plasticized polyacrylonitrile layer directly applied either to the electrode or the separator surfaces as an effective ion-conductive adhesive. Sanyo [174, 175] investigators, on the other hand, used thermally polymerizable additives to gel, or solidify, liquid electrolyte solutions in a wound, packaged battery.

The ceramic fillers (e.g., Al_2O_3, SiO_2, TiO_2) can greatly influence the characteristics and properties of polymer electrolyte by enhancing the mechanical stability and the conductivity [135, 175–178]. Prosini et al. [179] in a PVdF-HFP polymer matrix used γ-$LiAlO_2$, Al_2O_3, and MgO as fillers to form self-standing, intrinsically porous separators for lithium-ion batteries. The MgO-based separators showed the best anode and cathode compatibilities.

Liu et al. [180] has successfully prepared a PVdF-HFP/PE composite gel electrolyte by cast method. They showed that when the PE content was over 23 wt.%, the electrical impedance of the composite gel electrolyte increased rapidly by several orders, around the melting point of PE. The SEM pictures showed that the PE particles were fused and formed into a continuous film at or near the PE melting point, which cuts off the ion diffusion. This shutdown feature of the composite gel electrolyte can help in preventing the cell runaway under abusive usage. Similarly, Kim et al. [181] prepared polyethylene oxide (PEO)-coated separators by coating PEO onto a microporous PE separators. The ionic conductivity of PEO-coated membranes was higher than the base film. Kim et al. prepared the polymer electrolytes by coating polyethylene oxide (PEO) and polyethylene glycol dimethacrylate (PEGDMA) onto a microporous polyethylene membrane (Asahi Kasei, 25 um, 40% porosity) [182]. They showed that the relative weight ratio of PEO and PEGDMA coated onto the microporous membrane played a critical role in determining the uptake of electrolyte solution and ionic conductivity.

Separator for Aqueous Batteries

The aqueous batteries use water-based electrolytes (e.g., KOH electrolyte for NiCd, NiMH, and H_2SO_4 electrolyte for Lead acid), which are less resistive than nonaqueous electrolytes. Polyolefin materials are generally suitable for use in the manufacture of separators for these batteries, but they are not inherently wettable by aqueous electrolytes. Such electrolytes are therefore unable to penetrate the pores of a separator formed from such a material, so that ion migration through the pores in solution will not occur without modification. This problem is sometimes overcome by treating the polyolefin material with a surfactant, which allows an aqueous electrolyte to wet the material. However, such surfactant can be removed from the surfaces of the polyolefin material when electrolyte is lost from the device, for example during charging and discharging cycles, and it is not subsequently replaced on the material when the electrolyte is replenished.

This problem has also been addressed by modifying the surface properties of the polyolefin materials used to form polymeric sheets, by graft copolymerizing a monomeric substance to its surface, which, after copolymerization, confers hydrophilic properties, and, in some cases ion-exchange properties. This technique has been found to be practical when the porous substrate is formed from PE, which lends itself well to a graft-copolymerization reaction of this kind. However, it has been found that, when such a reaction is attempted using polyolefin materials other than PE, the rate of the grafting reaction is reduced significantly.

Graft polymerization is a convenient method for the modification of the physical and chemical properties of polymer materials and is of particular interest for synthesis of the hydrophilic membranes. Graft copolymerization can be achieved by various methods such as an exposure to ionizing radiation or ultraviolet light and the use of chemical initiators. Ionizing radiation is one of the most promising methods because of its rapid and uniform formation of active sites for initiating grafting throughout the matrix. Under appropriate experimental conditions, modifications of polymer properties can be accomplished not only on the surface but also throughout the polymer.

There have been several reports on radiation grafting of acrylic and methacrylic acid onto various substrates. These include both the direct grafting method and the pre-irradiation method to synthesize ion-exchange membranes. Two cation exchange membranes modified with the carboxylic acid group for battery separator was prepared by radiation-induced grafting of acrylic acid (AA) and methacrylic acid (MA) onto a polyethylene film by Choi et al. [183]. They found that KOH diffusion flux of AA-grafted PE membrane and MA-grafted PE membrane increased with an increase in the degree of grafting. AA-grafted PE membrane had a higher diffusion flux then MA-grafted PE membrane. Electrical resistance of both membranes decreased rapidly with an increase in the degree of grafting up to 120%, and then leveled off.

Battery separators having carboxylic acid group were prepared by radiation-induced grafting of acrylic acid onto a polyolefin nonwoven fabric (PNF). The PNF comprised of approximately 60% polyethylene and 40% polypropylene. It was found that the wetting speed, electrolyte retention, thickness, and ion-exchange capacity increased, whereas the electrical resistance decreased with increasing grafting yield [184]. The surface characteristics of the separators can also be modified by plasma discharge.

The subsequent subsections discuss separators used in lead acid and nickel metal hydride batteries.

Separators for Lead Acid Batteries

It has been a long time since the invention of the lead acid battery, but it still represents the most important secondary chemical power source – both in number of types and diversity of application. The lead acid battery has maintained its leading role for so many decades due to its competitive electrical characteristics and price, and due to its adaptability to new applications. It is manufactured in a variety of sizes and designs, ranging from less than 1 Ah to over 10,000 Ah [185].

Lead acid batteries can be classified into three major types or categories, namely, automotive (SLI), stationary, and motive power (industrial). In addition, there are many special batteries that cannot be easily categorized as either of the above types. These types of batteries are constructed with different materials and designed to meet the requirements of their intended end uses, each with a particular separator requirement with specific material composition, mechanical design, and physical, chemical, and electrochemical properties, tailored for the battery and its relevant specific uses. These batteries are generally available in flooded electrolyte or valve regulated (sealed) versions. In this section the types and properties of separators used for lead acid batteries are reviewed. The reader is referred to recent reviews published by Boehnstedt [12, 186, 187] and others [188–190] for detailed descriptions of lead acid separators.

Flooded Electrolyte Lead Acid Separators currently used in lead acid batteries can be classified based on their materials of construction into four major types: plastic (PE/silica, PVC/silica, Sintered PVC), paper (phenolic resin impregnated cellulose), glass (glass fiber mat), and rubber (hard rubber/silica, flexible rubber/silica, coated rubber/silica) separators. Table 8

Rechargeable Batteries, Separators for. Table 8 Typical separators used in lead acid battery systems

Separator	Class	Manufacturing process	Properties
Wood	Paper	Cellulosic separators are made from cotton linters or craft pulp and generally coated with phenolic resin for acid resistance and strength	Comparatively large pore size and relatively high electrical resistance
Hard rubber	Rubber	Made by mixing natural rubber, rehydrated precipitated silica, and sulfur. This is then extruded and calendared, vulcanized under water, and dried.	Finer pore diameter (0.2 μm average), relatively lower electrical resistance, excellent oxidation resistance, retards antimony transfer
Flexible rubber	Rubber	Made by mixing natural rubber, rehydrated precipitated silica. This is then extruded and calendared, irradiated with an ionizing electron beam and dried.	Flexible, fine pore structure (0.06 μm average), retards antimony transfer
Glass mat rubber	Rubber/ glass mat	Made by mixing polymeric emulsion, precipitated silica, and rubber. This is then coated on a fiberglass mat and finally cured and dried.	Finer pore diameter (<0.2 μm average), high porosity, excellent thermal dimensional stability
Sintered PVC	Plastic	Made by sintering PVC powder of a particle size ranging between 10 and 20 μm	Medium pore size (10–20 μm), generally good chemical resistance
Synthetic PVC	Plastic	Made from mixture of PVC, silica fine powder, and a solvent, and then extruded, calendared, and extracted	Small to medium average pore size and relatively low electrical resistance
Synthetic pulp with glass mat	Plastic/ glass	Made from blending PE synthetic pulp, synthetic fiber, and fine silica powder, and then heat treated	Medium pore size, low electrical resistance, and long service life at high temperatures; more difficult to process and assemble
Polyethylene (PE)	Plastic	Made from a mixture of UHMW PE powder, fine silica powder, and mineral oil. The mixture is extruded as a film, calendered, and made porous by extraction	Fine pore size, low electrical resistance, high puncture resistance, and strongly resistant to oxidation
Glass fiber mat	Glass	Deposition on a single sheet, a mixture of fibers dispersed in an aqueous solution	Excellent wettability, durable in an acid environment, good resiliency, high temperature stability, more difficult to process and assemble

shows the different types of separators used in batteries along with their manufacturing process and main features. Glass, paper, and sintered PVC separators can be classified as macroporous separators having an average pore diameter greater than 10 μm while all other separators can be classified as microporous separators having an average pore diameter smaller than 1 μm. All of these separators can be utilized as leaf separators in battery construction. Polyethylene can be used also as enveloped separators around either the positive or the negative plate. The use of "envelope" separators is popular in small, sealed cells, SLI, motive power, and standby batteries to facilitate production and to control lead contamination during manufacturing.

The environment of the lead acid battery (e.g., automotive battery) has been increasing in severity in recent years. The improvements and development of the separators have proceeded in accordance with the changes in the specifications for the batteries which were first made with wooden separators (preferred wood was Oregon Ceder as it contained small amounts of Lignin that enhances the performance of lead negative), then progressed through microporous rubber separators, cellulose separators and synthetic pulp separators (SPG) with glass mats, PVC separators, and now polyethylene separators have evolved. This sequential change in separator technology has provided continuous improvements in the charge and discharge efficiency of batteries and has given high vehicle-starting capability and reliability. Moreover, short circuits (caused by particles of active material dislodged from the battery plates) are prevented due to the smaller pores and excellent electrochemical oxidation resistance of the PE separator. These features contributed greatly to the improvement in battery life.

Rubber separators have good voltage characteristics, ability to retard antimony transfer, properties to retard dendrite growth, and good electrochemical compatibility [191]. Due to the hydrophilic properties of the rubber composition, the separators are highly wettable and renewable for the dry-charging process. Paik et al. showed that ACE-SIL (sulfur cured, hard rubber) separators performed well in industrial stationary or traction batteries, FLEX-SIL (electron-beam-cured, flexible rubber separator) separators are suited for deep-cycling batteries, and MICROPOR-SIL (a coated, glass mat, rubber separator) separators have been found to be a good choice for high rate discharging or cranking applications and for various types of gel cells [192]. Recently Daramic® DC UHMW PE has demonstrated excellent performance in these applications, as well.

Polyvinylchloride (PVC) and polyethylene (PE) separators have been the most commonly used separators in automotive batteries for the last 20 years. Polyethylene separators have a narrow pore-size distribution. The PVC separator is built up by sintering PVC powder in general of a particle size ranging between 10 and 20 μm. The decrease of particle size in the sintered product is negligible compared to the particle size of the raw materials. The pores are dispersed homogenously with a medium size ranging between 10 and 20 μm. Since a PVC separator exclusively consists of PVC, it exhibits advantageously good chemical resistance against acid and alkaline solutions. Unlike PE, PVC is disadvantaged due to its brittleness. The decline in PVC separators in recent years is in part due to their tendency to yield chloride ions from chemical attack.

The battery separators currently used by most flooded-cell-type lead acid battery manufacturers are of the microporous PE type. It was invented in the late 1960s by W.R. Grace & Co. [193]. The term "polyethylene separators" is somewhat misleading, since such a separator consists mainly of agglomerates of precipitated silica, being held within a network of extremely long-chain UHMWPE [194]. A typical PE separator formulation comprises precipitated silica (~60 wt.%), UHMW PE (~20 wt.%), mineral process oil (~15 wt.%), as well as some processing aids, like antioxidants and/or proprietary surface tension modifiers [195, 196].

The microporous PE separator is commercially manufactured by passing the ingredients through a heated extruder, passing the extrudate generated by the extruder through a die and into the nip formed by two heated calendar rolls to form a continuous web, extracting a substantial amount of the processing oil from the web by use of a solvent, drying the extracted web, slitting the web into lanes of predetermined width, and winding the lanes into rolls [195].

Rechargeable Batteries, Separators for. Table 9 Comparison of properties of different separators used in lead acid batteries

Property	Rubber	Cellulose	PVC	PE	Glass fiber
Year available	1930	1945	1950	1970	1980
Electrical resistance	Very poor	Poor	Poor	Very good	Very good
Porosity	Sufficient	Good	Poor	Good	Very good
Battery performance (cold crank)	Poor	Sufficient	Sufficient	Very good	Very good
Maximum pore size	Good	Poor	Sufficient	Very good	Poor
Mean pore diameter	Good	Poor	Poor	Very good	Poor
Purity	Good	Fair	Good	Good	Good
Resistance to shorting	Good	Poor	Poor	Very good	Poor
Corrosion resistance	Very good	Poor	Good	Very good	Good
Oxidation resistance	Fair	Poor	Good	Very good	Very good
Envelopable (sealability)	Very poor	Very poor	Sufficient	Very good	Very poor
Flexibility	Brittle	Brittle	Brittle	Excellent	Good

The PE separators have excellent microporous structure for electrolyte flow with minimal lead particle deposits; excellent ductility, strength, and toughness for envelopability and plate puncture resistance; excellent oxidation, chemical, and thermal resistance to resist premature deterioration; good manufacturability with high production efficiency and relatively low raw material cost which reduces overall manufacturing costs [196]. The PE pocket separation is in almost all aspects significantly superior to leaf separation. Only PE separators can be enveloped and can develop good sealability. These have low electrical resistance, sufficient porosity, small pore size, and great resistance to both shorting and corrosion. The PE separator, by virtue of its low electrical resistance, generally provides better cold cranking performance. They are very flexible and offer excellent oxidation resistance if the residual oil content is controlled and/or proprietary chemical modifiers have been incorporated. A comparison of the properties of different types of separators is given in Table 9.

PE separators have contributed to improved battery specific energy and specific power, increased battery cycle life, and higher temperature operating capabilities. PE separators have gained in popularity and have generally replaced PVC, cellulose, glass fiber, and other conventional separators. The transition to microporous PE envelope separators started in the USA in the 1970s, followed by Europe in 1980s. Today, PE separators have captured almost 100% of the US market and more than 70% of the remaining worldwide automotive markets [197].

In a flooded-cell-type lead acid battery, the battery separator typically has “ribs” or protrusions extending from at least one planer face of the separator. Such ribs are either formed integrally with the backweb of the separator, or they can be subsequently applied to the backweb as a bead of the same or different material as the backweb, or they can be formed by embossing the backweb. The ribs function is to provide proper spacing between the plates and to provide a space wherein free electrolyte resides. The ribs also provide pressure to hold the electrodes in contact with the separator. This reduces the need for precise dimensional control on the cell components. Microporous PE separators typically have a configuration comprising a backweb having a predetermined thickness, and a plurality of parallel or patterned ribs spaced apart by a predetermined distance and extending outwardly from one planar surface of the backweb. The ribs extend continuously in a longitudinal direction parallel to the edges of the separator material. The thickness of the backweb and height and spacing of the ribs is specified to the separator manufacturer by the battery manufacturer; based on specifications designed to maximize certain battery characteristics desired by the battery manufacturer. SLI batteries tend to have separators that are thinner than "industrial" lead acid batteries used for standby power sources and traction devices.

Endoh [198] has reported that one reason for the occasionally shortened life of batteries assembled with PE pocket separators is the development of internal short circuits at the bottom part of the PE separator due to anodic corrosion causing active material to shed from the positive plates and leak through the separators. He also found that when synthetic pulp (SP) separators with glass mats are used, it is possible to not only restrain the shedding from positive plates, but also to protect the separators against intensive oxidation so that no internal short circuits develop on charge. He concluded that the use of SP separators with glass mat is required to produce long service-life batteries, especially in tropical regions.

Higashi et al. [199] carried out endurance test under high temperature conditions on automotive batteries made with three different types of separators. One group was assembled with PE pocket separators for the negative plates, another with PE pocket separators with glass mats for the positive plates, and a third with leaf-type synthetic pulp separators with glass mats. They concluded that battery assembly with PE pocket separators with glass mat is an effective way to achieve good endurance (i.e., life extension at high temperature) and leaf-type synthetic pulp separators with glass mats are the best approach for hot climatic conditions.

Valve Regulated Lead Acid (VRLA) The valve regulated lead acid battery is an important development in lead acid battery technology. These batteries operate on the principle of oxygen recombination, using a "starved" or immobilized electrolyte. The oxygen generated at the positive electrode during charge can, in these battery designs, diffuse to the negative electrode, where it can react, in the presence of sulfuric acid, with freshly formed lead. The separator material should provide innumerable gas channels between the plates through which oxygen can flow from the positive to the negative electrode. These batteries differ from its flooded electrolyte precursor in a number of important ways [200]. They have been manufactured for many years with microfiber glass separators, also called absorptive glass mat (AGM). They are inherently resistant to acid stratification and have the additional important advantage of being essentially maintenance free. The separator is a crucial component in determining the useful life of a VRLA cell. While a prime function of the VRLA cell separator is to hold the cell's electrolyte in place, it must also offer characteristics that prevent major failure mechanisms occurring in the cell's positive and negative plates.

The microglass separator, since its discovery by McClelland and Devitt, has been the material of choice for VRLA designs [201, 202]. It is a wet laid nonwoven (glass fiber) "paper" and is manufactured on a paper machine. The type of paper machine used by the manufacturer can influence the separator properties. Three properties – porosity, uniformity, fiber directionality – are important attributes that can be influenced by the type of fiber used. The glass fiber, which has a zero contact angle with the acid, is durable in the acid environment, and the fine fiber structure also has good resiliency to allow for a sustained pressure against the plate. The microglass separator has a porosity in the 90–95% range and is very conformable. It can adapt to imperfections in the plate surface. The separator also has high temperature stability. Recent studies have shown that higher levels of fine fiber and higher separator compression provide improved cycle performance in VRLA batteries [203–206].

On the other hand, AGM separators offer little control over the oxygen transport rate or the recombination process. The arrival of too much oxygen to the negative plate could result in overheating, hindrance of the battery's ability to recharge, or even a loss of capacity. Furthermore, AGM separators exhibit low puncture resistance.

Nickel Systems

The nickel-based systems have traditionally included the following systems – nickel-iron (Ni/Fe), nickel-cadmium (NiCd), nickel metal hydrides (NiMH), nickel hydrogen (Ni/H_2), and nickel-zinc (Ni/Zn). Of these, the metal hydride chemistry has been the most successful in the secondary battery market. All nickel systems are based on the use of a nickel oxide active material (undergoing one valence change from charge to discharge or vice-versa). The electrodes can be pocket type, sintered type, fibrous type, foam type, pasted type, or plastic roll-bonded type. All systems use an alkaline electrolyte, KOH.

The sealed nickel metal hydride battery uses hydrogen, absorbed in a metal alloy, for the active negative material. The NiMH batteries have a higher energy density and are considered environmentally friendly than the NiCd battery. However, the sealed NiMH battery has limited rate capability and is less tolerant of overcharge. The self-discharge rate is generally higher when conventional nylon separators are used [207]. The presence of oxygen and hydrogen gases cause the polyamide materials to decompose, producing corrosion products which poison the nickel hydroxide, promoting premature oxygen evolution and also forming compounds capable of a redox shuttle between the two electrodes which further increases the rate of self-discharge [208]. Ikoma et al. carried out a detailed investigation to study the self-discharge mechanism and contribution of separators [209]. They used nonwoven fabric made of conventional polyamide (PA), PP (with surfactant), and a nonwoven fabric whose main material was sulfonated-PP (hydrophilic) as separators. When nonwoven fabric made of chemically stable sulfonated-PP is used as a separator instead of a conventional polyamide separator, the self-discharge rate of the NiMH battery was strongly depressed, to the same level as that for the NiCd battery [208, 210, 211].

Nagarajan et al. [212] used differential scanning calorimeter (DSC) to study the materials used as separators in commercial AA cells. They found that Sanyo and Matsushita cells containing nonwoven fabrics fabricated from conventional polyamide as separators exhibited substantially higher self-discharge due to the shuttle reaction of the ammonia and amine. Scimat Ltd. has shown that acrylic acid grafted nonwoven polyolefin separators have the ability to absorb chemical impurities, for example ammonia, from the alkaline environments. It has been shown that by using a grafted polyolefin separator the free ammonia present inside a NiMH cell is trapped by the separator resulting in a reduction in self-discharge to levels normally associated with NiCd cells [213]. In October 2002, Scimat Ltd. announced the launch of the next generation of separators for NiMH and NiCd cells using its second-generation grafting technology [214].

The commonly used separator material now is the surface-treated polypropylene. The surface treatment helps in making the polypropylene permanently wettable. Surface treatments involve the grafting of a chemical such as acrylic acid to the base fibers to impart wettability and are accomplished using a variety of techniques such as UV or cobalt radiation. Another method of imparting wettability to the polypropylene is a sulfonation treatment where the base fiber material is exposed to fuming sulfuric acid. The separator surface is designed to be made hydrophilic to the electrolyte.

Cheng et al. [215] carried out the impedance study on a foam-type NiMH battery with nonwoven PP separator to determine the main causes of early cycle deterioration. Their data indicated that the decrease in the voltage characteristic of the battery was due to drying out of the separator that increases the ohmic resistance of the battery, and that decay of the total discharge capacity is due to an inactive surface that increases the charge-transfer resistance of the battery.

Mathematical Modeling of Separators

Computer simulations have been used as an important tool for understanding and optimizing battery performance since the 1970s [216–218]. Continued progress in computational tools has enabled ever-increasing sophistication in battery modeling and a steady increase in the number of systems to which modeling has been applied. Today, it is possible to obtain simulation codes for all of the major rechargeable batteries, some of which are available in the public domain [219].

The mathematical models of different types of batteries (lead acid [220, 221], NiMH [222], lithium-ion [223, 224]) have been developed during the last few years [219]. This has led to a better understanding of those systems. The present models consider usually, the thickness and porosity of the separators. Very little has been done in incorporating the effect of physical and chemical properties of separators on the performance and safety of batteries. This is also because the microstructure of separators and their effect on transport properties in batteries are generally known only qualitatively.

A thorough understanding of the microstructure of separators would be beneficial for modeling studies and optimization of electrochemical systems. This will help in making the battery model predictions more practical and reliable. The separator pore structure is usually very complex. It consists of a porous network of interconnected pores, which are filled with liquid electrolyte. A complete description of the pore structure would require a very intricate model. Simulations are only practically possible if a simplified quasi-continuum model involving a few parameters represents the structure. In such an approach, the "effective" electrolytic conductivity, σ_{eff}, is often defined by [93]

$$\sigma_{\mathrm{eff}} = \varepsilon^{\alpha}\sigma_0, \alpha \approx 1.5 \tag{11}$$

where σ_0 is the bulk ionic conductivity of the electrolyte, ε is the void volume fraction of separator filled with electrolyte, and α is the Bruggeman exponent. The general applicability of alpha ∼1.5 appears questionable because separator pores are never of an ideal shape. Fan and White [224] chose a α value of 2.5 for separators in NiCd batteries and Doyle et al. [225] used 3.3 for lithium-ion batteries. Arora et al. [226] measured the value as 2.4 for PVdF-based separators by measuring the separator and electrolyte conductivity at different salt concentration. Doyle et al. used an even higher Bruggeman exponent of 4.5 for quantifying the ionic conductivity of their plasticized electrolyte membrane [86].

Patel et al. showed that a Bruggeman exponent of 1.5 is often not valid for real separator materials, which do not have uniform spherical shape [227]. Porous networks based on other morphologies such as oblate (disk-type) ellipsoids or lamella increase the tortuous path for ionic conductivity and result either in a significant increase of the exponent α, or in a complete deviation from the power law. They showed that spherical or slightly prolate ellipsoidal pores should be preferred for separators, as they lead to higher ionic conductivity separators.

Tye [122] explained that separator tortuosity is a key property determining transient response of a separator and steady-state electrical measurements do not reflect the influence of tortuosity. He recommended that the distribution of tortuosity in separators be considered; some pores may have less tortuous paths than others. He showed mathematically that separators with identical average tortuosity and porosities could be distinguished by their non-steady-state behavior if they have different distributions of tortuosity.

Doyle et al. [86] used a mathematical model to examine the effect of separator thickness for the PVDF:HFP gel electrolyte system and found that decreasing separator thickness below 52 μm caused only a minor decrease in ohmic drop across the cell. The voltage drop in the electrodes was much more significant. Mao and White [228] developed a mathematical model for discharge of a Li/TiS$_2$ cell. Their model predicted that increasing the thickness of the separator from 25 to 100 μm decreased the discharge capacity from 95% to about 90%; further increasing separator thickness to 200 μm reduced discharge capacity to 75%. These theoretical results indicate that conventional separators (25–37-μm thick) do not significantly limit mass transfer of lithium. Santhanagopalan et al. [229] studied the influence of the separator on several internal short-circuit scenarios.

The use of electroactive polymers for overcharge protection has been recently reported for lithium-ion batteries [230, 231]. The electroactive polymer incorporated into a battery's separator is an attractive new option for overcharge protection. Thomas et al. [233] developed a mathematical model to explain how electroactive polymers such as polythiophene can be used to provide overcharge protection for lithium-ion batteries. The model shows that, as the cell potential exceeds the oxidation potential of the polymer, the cell is transformed, over a time scale of a few minutes, from a battery into a resistor, after which a steady-state overcharge condition is attained.

Summary

The ideal battery separator would be infinitesimally thin, offer no resistance to ionic transport in electrolytes, provide infinite resistance to electronic conductivity for isolation of electrodes, be highly tortuous to prevent dendritic growths, and be inert to chemical reactions. Unfortunately, such a product is not commercially feasible. Actual separators are electronically insulating membranes whose ionic resistivity is brought to the desired range by manipulating the membranes thickness and porosity.

It is clear that no single separator satisfies all the needs of battery designers, and often, optimization is performed on a case-by-case basis. It is ultimately the application that decides which separator is most suitable. This chapter is intended to be a useful tool to help battery manufacturers in selecting the most appropriate separators for their batteries and respective applications. The information provided is purely technical and does not include other very important parameters, such as cost of production, availability, etc.

There has been a continued demand for thinner battery separators to increase battery power and capacity. This has been especially true for lithium-ion batteries used in portable electronics. However, it is very important to ensure the continued safety of batteries, and this is where the role of the separator is greatest. Thus, it is essential to optimize all the components of battery to improve the performance while maintaining the safety of these cells. Separator manufacturers continue to work along with the battery manufacturers to create the next generation of batteries with increased reliability and performance, but always keeping safety in mind.

This chapter has attempted to present a comprehensive review of literature on separators used in various batteries. It is evident that a wide variety of separators are available, and that they are critical components in batteries. In many cases, the separator is one of the major factors limiting the life and/or performance of batteries. Consequently, development of new improved separators would be very beneficial for the advanced high capacity batteries.

Future Directions

Up until the last few years, most of the separators and membranes historically used had not been customized for specific battery applications. Thus, future research should be aimed at developing custom-separators that are specifically tailored for the individual battery applications. For example, the form factor of batteries has drastically changed since 2005 – with the advent of sleek consumer electronic devices, and large format batteries and packs for automotive applications. Efforts toward better heat dissipation, extended operating windows, and more stringent requirements on long-term durability have significantly expanded. The general objectives of separator research should be: (a) to find new and cost-effective separators, (b) to understand the separator properties in batteries, and (c) to optimize separator properties related to specific cell performance, life, and safety. The battery separators for tomorrow will demand more than just good insulation and mechanical filtration; they will require unique electrochemical properties. More work focusing on precise control of the membrane properties guided by fundamental insight into the membrane behavior will set the trend for the next generation separators.

Acknowledgments

The authors would like to acknowledge numerous colleagues and staff members at Celgard, LLC and Polypore, International Inc. for their help with the preparation and publication of this chapter.

Bibliography

Primary Literature

1. Linden D, Reddy TB (2002) Handbook of batteries, 3rd edn. McGraw Hill, New York
2. Besenhard JO (ed) (1999) Handbook of battery materials. Weimheim, Wiley-VCH
3. Berndt D (2003) Maintenance free batteries, 3rd edn. Research Studies Press, Taunton, Somerset
4. Arora P, Zhang Z (2004) Chem Rev 104:4419
5. Brodd RJ, Friend HM, Nardi JC (eds) (1995) Lithium ion battery technology. ITE-JEC Press, Brunswick
6. Wakihara M, Yamamoto O (eds) (1998) Lithium ion batteries, fundamentals and performance. Wiley-VCH, New York
7. Yoshino A (1995) Chem Ind 146:870

8. Schalkwijk WAV (ed) (2002) Advances in lithium ion batteries. Kluwer, New York
9. Kinoshita K, Yeo R (1985) Survey on separators for electrochemical systems. LBNL, January 1985
10. Zhang SS (2007) J Power Sources 164 (1):351
11. Benett J, Choi WM (1995) Developments in small cell separators. In: Proceedings of the 10th annual battery conference on applications & advances, IEEE, New York, 10–13 Jan 1995, p 265
12. Boehnstedt W (1999) In: Besenhard JO (ed) Handbook of battery materials. VCH Wiley, Amsterdam
13. Spotnitz R (1999) In: Besenhard JO (ed) Handbook of battery materials. VCH Wiley, Amsterdam
14. Shirai H, Spotnitz R (1996) In: Kozawa A, Yoshio M (eds) Lithium ion secondary battery – materials and applications. Nikkan Kogyo Shin-bun, Tokyo, p 91 (in Japanese)
15. Shirai H, Spotnitz R, Atsushi A (1997) Characterization of separators for lithium ion batteries – a review. Chem Ind 48:47 (in Japanese)
16. Hiroshi T, In: Ogumi Z (ed) (1997) The latest technologies of the new secondary battery materials. CMC, Tokyo, p 99 (in Japanese)
17. Hiroshi T, In: Oyama N (ed) (1998) Advanced technologies for polymer battery. CMC, Tokyo, p 165 (in Japanese)
18. Koichi K, In: Oyama N (ed) (1998) Advanced technologies for polymer battery. CMC, Tokyo, p 174 (in Japanese)
19. Kiyoshi K, In: Kanamura K (ed) (2004) Lithium secondary battery technology for the 21st century. CMC, Tokyo, p 116
20. Brodd RJ, Bullock KR, Leising RA, Middaugh RL, Miller JR, Takeuchi EJ (2004) Electrochem Soc 151:K1
21. http://www.freedoniagroup.com/Batteries.html. Accessed Feb 2010
22. About Edison batteries, Inc. http://www.optodot.com/systmpl/htmlpage/. Accessed Feb 2010
23. Industry News: Looking Back, Looking Forward. Battery & EV Technology, January 2004, 28:2
24. Advanced rechargeable battery industry 2001/2002. Nomura Research Institute Limited, Japan
25. Battery and fuel cell components (2009) The Fredonia Group, Cleveland, OH
26. Pilot C (2004) The worldwide rechargeable battery market. Presented at Batteries 2004, 6th edn, Paris, 2–4 June 2004
27. Celgard Inc. http://www.celgard.com. Accessed Feb 2010
28. http://www.celgard.com/news-room/press-releases/2010/celgard-expansion.asp. Accessed Feb 2010
29. http://www.asahi-kasei.co.jp/asahi/en/ir/supplement/1003transcript.pdf. Accessed Feb 2010
30. Advanced Battery Technology, February 2004, 40:22
31. Pekala RW, Khavari M (2003) US Patent 6,586,138
32. www.porouspower.com. Accessed Feb 2010
33. http://www.sumitomo-chem.co.jp/english/division/itrc.html. Accessed Feb 2010
34. Hamano K, Yoshida Y, Shiota H, Shiraga S. Aihara S, Murai M, Inuzuka T (2003) US Patent 6,664,007 B2
35. Sun L (2003) US Patent 2003/0152828A1
36. Johnson BA, White RE (1998) J Power Sources 70:48
37. Hoffman HG (1995) In: Proceedings of the tenth annual battery conference on applications and advances, LongBeach, 10–13 Jan 1995, IEEE, New York, p 253
38. Geiger M, Callahan RW, Diwiggins CF, Fisher HM, Hoffman DK, Yu WC, Abraham KM, Jillson MH, Nguyen TH (1994) The eleventh international seminar on primary and secondary battery technology and application, Fort Lauderdale, FL, 28 Feb–3 Mar 1994, Florida Educational Seminars, Boca Raton
39. Tanba H (1999) Molding Process 11:759
40. Adachi A, Spotnitz RM et al (1997) Osaka chemical marketing center. Osaka, Japan, p 69
41. Callahan RW, Nguyen KV, McLean JG, Propost J, Hoffman DK (1993) In: Proceedings of the 10th international seminar on primary and secondary battery technology and application, Fort Lauderdale, 1–4 Mar 1993. Florida Educational Seminars, Boca Raton
42. Yu WC, Hux SE (1999) US Patent 5,952,120
43. Hipore. http://www.asahikasai.co.jp/memrbane/english/tradenm/t07.html. Accessed Feb 2010
44. Druin ML, Loft JT, Plovan SG (1974) US Patent 3,801,404
45. Schell WJ, Zhang Z (1999) In: The fourteenth annual battery conference on applications and advances, Long Beach, 12–15 Jan 1999. IEEE, New York, p 161
46. Isaacson RB, Bierenbaum HS (1971) US Patent 3,558,764
47. Kamei E, Shimomura Y (1986) US Patent 4,563,317
48. Yu TH (1996) Processing and structure-property behavior of microporous polyethylene – from resin to final film. Ph.D. Dissertation, Virginia Polytechnic Institute and State University, Virginia
49. Sarada T, Sawyer LC, Ostler MI (1983) J Membr Sci 15:97
50. Bierenbaum HS, Daley LR, Zimmerman D, Hay IL (1974) US Patent 3,843,761
51. Hamer EAG (1986) US Patent 4,620,956
52. Hiroshi K, Tetuo A, Akira K (1997) US Patent 5,691,047
53. Kesting RE (1985) Synthetic polymeric membranes, 2nd edn. Wiley, New York
54. Ihm DW, Noh JG, Kim JY (2002) J Power Sources 109:388
55. Bierenbaum HS, Isaacson RB, Druin ML, Plovan SG (1974) Ind Eng Chem Prod Res Dev 13:2
56. Norimitsu K, Kotaru T, Koichi K, Hidehiko F (2000) US Patent 6,153,133
57. Michiyuki A (1996) Japan Patent 8064194
58. Kotaro T, Koichi K, Tatsuya T, Kenkichi O (1991) US Patent 5,051,183
59. Koichi K, Kotaro T, Mamoru T, Tatsuya T (1996) Japan Patent 8012799
60. Akinao H, Kazuo Y, Hitoshi M (2000) US Patent 6,048,607
61. Userguide (2003) FreedomCar separator costing document, February 2003
62. Xu M, Hu S, Guan J, Sun X, Wu W, Zhu W, Zhang X, Ma Z, Han Q, Liu S (1992) US Patent 5,134,174

63. Fisher HM, Wensley CG (2002) US Patent 6,368,742
64. Zhu W, Zhang X, Zhao C, Zu W, Hou J, Xu M (1996) Polym Adv Tech 7:743
65. Sadamitsu K, Ikeda N, Hoki M, Nagata K, Ogino K (2002) World Patent Application 02066233A1
66. Higuchi H, Matsushita K, Ezoe M, Shinomura T (1995) US Patent 5,385,777
67. Calis GHM, Daemen APM, Gerrits NSJA, Smedinga JT (1997) J Power Sources 65:275
68. Ooms FGB, Kelder EM, Schoonman J, Gerrits N, Smedinga J, Calis G (2001) J Power Sources 97–98:598
69. Yamamura Y, Ooizumi S, Yamamoto K (2001) Separator for rechargeable lithium-ion batteries with high puncture strength and high melt rupture temperature, Nitto Denko Technical Report, vol 39, p 39. http://www.nitto.com/rd/rd6_1.html. Accessed Feb 2010
70. Pekala RW, Khavari M, Dobbie G, Lee D, Fraser-Bell G (2000) In: 17th international seminar & exhibit on primary and secondary batteries, Fort Lauderdale, 6–9 Mar 2000. Florida Educational Seminars, Boca Raton
71. Fleming R, Taskier H (1990) Prog Batt Sol Cells 9:58
72. Hoffman D, Fisher H, Langford E, Diwiggins C (1990) Prog Batt Sol Cells 9:48
73. Spotnitz R, Ferebee M, Callahan RW, Nguyen K, Yu WC, Geiger M, Dwiggens C, Fischer H, Hoffman D (1995) In: Proceedings of the 12th international seminar on primary and secondary battery technology and applications, Fort Lauderdale, 6–9 Mar 1995. Florida Educational Seminars, Boca Raton
74. Yu WC, Callahan RW, Diwiggins CF, Fischer HM, Geiger MW, Schell WJ (1994) In: North America membrane society conference, Breckenridge, CO
75. Kuribayashi I (1996) J Power Sources 63:87
76. Pasquier AD, Gozdz A, Plitz I, Shelburne J (2002) In: 201st meeting, The Electrochemical Society, Philadelphia, 12–17 May 2002
77. Augustin S, Volker H, Gerhard H, Christian H (2002) Desalination 146:23
78. http://www.separion.com. Accessed Feb 2010
79. Hying C (2004) Separion separators for lithium batteries – safety & performance. In: Batteries 2004, 6th edn. Paris, 2–4 June 2004
80. Sachan S, Ray CA, Perusich SA (2002) Polym Eng Sci 42:1469
81. Sachan S, Perusich S (1999) In: Electrochemical Society Meeting, Seattle
82. Carlson SA (2004) Membrane & separation technology news. 22, 8
83. Abraham KM (1993) Electrochim Acta 38:1233
84. Gineste JL, Pourcell G (1995) J Membr Sci 107:155
85. Hoffman DK, Abraham KM (1991) In: Proceedings of the fifth international seminar on lithium battery technology and applications, Deerfield Beach. Florida Educational Seminars, Boca Raton, FL
86. Fuller TF, Doyle M, Newman J (1994) J Electrochem Soc 141:1
87. USABC (2001) Development of low cost separators for lithium-ion batteries, RFPI 2001
88. Laman FC, Sakutai Y, Hirai T, Yamaki J, Tobishima S (1992) Extended abstract. In: 6th International meeting lithium batteries, Münster, Germany, 10–15 May 1992, p 298
89. Laman FC, Gee MA, Denovan J (1993) J Electrochem Soc 140:L51
90. Robinson RG, Walker RL (1963) In: Collins DH (ed) Batteries. MacMillan, New York, p 15
91. Lander JJ, Weaver RD, Salkind AJ, Kelley JJ (1964) In: Cooper JE, Fleischer A (eds) Characteristics of separators for alkaline silver oxide zinc secondary batteries. Screening methods. NASA Technical Report NAS 5–2860
92. Kilroy WP, Moynihan CT (1978) J Electrochem Soc 125:520
93. MacMullin RB, Muccini GA (1956) AIChE J 2:393
94. Spotnitz R, Ferebee MW (1996) Meeting abstracts, 6–11 Oct 1996. The Electrochemical Society Inc., vol 96-2, Fall Meeting, San Antonio, TX
95. Ionov VV, Isakevitch VV, Katalevsky EE, Chernokoz AJ (1990) J Power Sources 30:321
96. Lowell S, Shields E (1991) Powder surface area and porosity, 3rd edn. Chapman and Hill, New York
97. PMI conference 2000 proceedings (2000) PMI short course, Ithaca, NY, 16–19 Oct 2000
98. Porous materials Inc. http://www.pmiapp.com. Accessed Feb 2010
99. Jena AK, Gupta KM (1999) J Power Sources 80:46
100. Jena AK, Gupta KM (2001) J Power Sources 96:214
101. Venugopal G, Moore J, Howard J, Pendalwar S (1999) J Power Sources 77:34
102. Zeman L, Denault L (1992) J Membr Sci 71:221
103. Chen RT, Saw CK, Jamieson MG, Aversa TR, Callahan RW (1994) J Appl Polym Sci 53:471
104. Fujii T, Mochizuki T (1998) US Patent, 5,759,678
105. Venugopal G (1997) The role of plastics in lithium-ion batteries. In: Proceedings of the 3rd annual conference on plastics for portable and wireless electronics, Philadelphia, 14–15 Oct 1997, p 11
106. Lundquist JT, Lundsager CB, Palmer NL, Troffkin HJ, Howard J (1998) US Patent 4,731,304
107. Lundquist JT, Lundsager CB, Palmer NI, Troffkin HJ (1987) US Patents 4,650,730
108. Faust MA Suchanski MR, Osterhoudt HW (1988) US Patent No. 4,741,979
109. Matthias U, Dieter B, Heinrich R, Thomas B (2003) US Patent 6,511,517
110. Maleki H, Shamsuri AK (2003) J Power Sources 115:131
111. http://www.cpsc.gov/cpscpub/prerel/prhtml07/07011.html. Accessed Feb 2010
112. Norin L, Kostecki R, McLarnon F (2002) Electrochem Solid State Lett 5:A67
113. Kostecki R, Norin L, Song X, McLarnon F (2004) J Electrochem Soc 151:A522

114. Hazardous materials regulations, Code of Federal Regulations, CFR49 173.185
115. UL1640, Lithium batteries, Underwriters Laboratories, Illinois
116. UL2054, Household and commercial batteries, Underwriter Laboratories, Illinois
117. Secondary lithium cells and batteries for portable applications, International Electrotechnic Commission, IEC 61960–1 and IEC 61960–2
118. Recommendations on the transport of dangerous goods (1999) Manual of Tests and Criteria, United Nations, New York
119. Safety Standard for Lithium Batteries (1995) UL 1642, 3rd edn. Underwriters Laboratories Inc, Illinois
120. Standard for Household and Commercial Batteries (1993) UL 2054. Underwriter Laboratories, Illinois
121. UN Recommendations on the Transport of Dangerous Goods, December 2000
122. Tye FL (1983) J Power Sources 9:89
123. Japan Battery Association (1997) A guideline for the safety evaluation of secondary lithium cells. Japan Battery Association, Tokyo
124. Venugopal G (2001) J Power Sources 101:231
125. Kim MH (2008) LG Li-Ion technology for automotive applications. Presented at the AABC 2008, Tampa, FL
126. Alamgir M, Abraham KM (1994) Chapter 3. In: Pistoia G (ed) Lithium batteries: new materials, developments and perspectives, vol 5, Industrial Chemistry Library. Elsevier, New York
127. Gray FM (1997) Polymer electrolytes, RSC materials monograph. The Royal Society of Chemistry, Cambridge
128. Fauteux D, Massucco A, McLin M, Van Buren M, Shi J (1995) Electrochim Acta 40:2185
129. North M, Markin TL, Hooper A, Tofield BC (1984) In: Second international meeting on lithium batteries, Extended Abstracts #19, Paris, France, 25–27 Apr 1984
130. Appetecchi GB, Dautzenberg G, Scrosati BJ (1996) J Electrochem Soc 143:6
131. Armand M (1983) Solid State Ionics 9/10:745
132. Lightfoot P, Mehta MA, Bruce PG (1993) Science 262:883
133. Vincent CA, Scrosati B (1993) Modern batteries. An introduction to electrochemical power sources. Arnold, London
134. Appetecchi GB, Passerini S (2002) J Electrochem Soc 149:A891
135. Croce F, Appetecchi GB, Persi L, Scrosati B (1998) Nature 394:4496
136. Croce F, Persi L, Ronci F, Scrosati B (2000) Solid State Ionics 135:47
137. Fan J, Fedkiw PS (1997) J Electrochem Soc 144:399
138. Appetecchi GB, Romagnoli P, Scrosati B (2001) Electrochem Commun 3:281
139. Kim DW, Sun YK (2001) J Power Sources 102:41
140. Chojnacka J, Acosta JL, Morales E (2001) J Power Sources 97–98:819
141. Nishi Y (2002) In: van Schalkwijk W, Scrosati B (eds) Advances in lithium ion batteries. Kluwer/Plenum, New York
142. EE Times.com. http://www.eet.com/story/OEG19990121S0013. Accessed Feb 2010
143. Song JY, Wang YY, Wan CC (1999) J Power Sources 77:183
144. Min HS, Ko JM, Kim DW (2003) J Power Sources 119–121:461
145. Jo SI, Sohn HJ, Kang DW, Kim DW (2003) J Power Sources 119–121:478
146. Kim HS, Kum KS, Cho WI, Cho BW, Rhee HW (2003) J Power Sources 124:221
147. Abraham KM, Alamgir M (1994) Solid State Ionics 70–71:20
148. Schmutz C, Tarascon JM, Gozdz AS, Schumtz CN, Warren PC, Shokoohi FK (1995) In: Proceeding electrochemical society, 1995. Rechargeable lithium and lithium ion batteries, vol 94–28, pp 330–335
149. Murata K, Izuchi S, Yoshihisa Y (2000) Electrochim Acta 45:1501
150. Jiang Z, Carroll B, Abraham KM (1997) Electrochim Acta 42:2667
151. Song JY, Cheng CL, Wang YY, Wan CC (2002) J Electrochem Soc 149:A1230
152. Gozdz AS, Schmutz CN, Tarascon JM, Warren PC (1995) US Patent 5,456,000
153. Gozdz AS, Tarascon JM, Schmutz CN, Warren PC, Gebizlioglu OS, Shokoohi F (1995) In: Tenth annual battery conference on advances and applications, Long Beach, 10–13 Jan 1995. IEEE, New York, p 301
154. Tarascon JM, Gozdz AS, Schumtz CN, Shokoohi FK, Warren PC (1996) Solid State Ionics 86–88:49
155. Pasquier AD, Waren PC, Culver D, Gozdz AS, Amatucci G, Tarascon JM (1999) Electrochem Soc Proc 99(24):360
156. Park CK, Kakirde A, Ebner W, Manivannan V, Chai C, Ihm DJ, Shim YJ (2001) J Power Sources 97–98:775
157. Dasgupta S, Jacobs JK (1995) US Patent 5,437,692
158. Abraham KM, Alamgir M, Hoffman DK (1995) J Electrochem Soc 142:683
159. Pendalwar SL, Howard JN, Venugopal G, Oliver M (1998) US Patent 5,716,421
160. Gozdz AS, Plitz I, Du Pasquier A, Zheng T (2001) In: Proceedings of the 200th ECS meeting, Phoenix, AZ, vol 2000–2001, Fall 2001, p 336
161. Gozdz AS (2001) US Patent 6,328,770
162. Kim DW, Oh B, Park JH, Sun YK (2000) Solid State Ionics 138:41
163. Wang Y, Sejdic JT, Steiner R (2002) Solid State Ionics 148:443
164. Gozdz AS, Plitz I, DuPasquier A, Zheng T (2007). Presented at the 198th meeting of the electrochemical society, Phoenix, AZ, 22–27 Oct 2000
165. Spotnitz RM, Wensley CG (2000) US Patent 6,322,923
166. Fabrice C, Bradford R (2002) WO 02/50929 A2
167. Jeong YB, Kim DW (2004) J Power Sources 128:256
168. Dasgupta S, Jacobs JK (1996) US Patent 5,498,489
169. Balsara N (2010) Block copolymer separators for lithium batteries. In: Presented at the 2010 annual merit review of the vehicle technologies program. http://www1.eere.energy.gov/vehiclesandfuels/pdfs/merit_review_2010/electrochemical_storage/es088_balsara_2010_p.pdf. Accessed Feb 2010
170. Eschbach FO, Oliver M (1997) US Patent 5,681,357
171. Hamano K, Shiota H, Shiraga S, Aihara S, Yoshida, Y, Murai M, Inuzuka T (1999) US Patent 5,981,107
172. Akashi H (1997) US Patent 5,658,686

173. Akashi H (1996) Paper presented at the international symposium on polymer electrolytes, ISPE-5, Uppsala, Sweden, 11–16 Aug 1996
174. Fujii T (2000) In: Proceeding of 17th international seminar & exhibit on primary and secondary batteries, Boca Raton, 6–9 Mar 2000. Florida Educational Seminars, Boca Raton
175. Nakane I, Narukawa S (2000) Power 2000. In: The 8th annual international conference on power requirements for mobile computing and wireless communications, San Diego, CA, 24–27 Sept 2000
176. Kim KM, Ryu KS, Kang SG, Chang SH, Chung IJ (2001) Macromol Chem Phys 202:866
177. Kim KM, Park NG, Ryu KS, Chang SH (2002) Polymer 43:3951
178. Kim KM, Ko JM, Park NG, Ryu KS, Chang SH (2003) Solid State Ionics 161:121
179. Prosini PP, Villano P, Carewska M (2002) Electrochim Acta 48:227
180. Liu X, Kusawake H, Kuwajima S (2001) J Power Sources 97–98:661
181. Kim DW, Ko JM, Chun JH, Kim SH, Park JK (2001) Electrochem Comm 3:535
182. Kim DW, Noh A, Chun JH, Kim SH, Ko JM (2001) Solid State Ionics 144:329
183. Choi SH, Park SY, Nho YC (2000) Radiat Phys Chem 57:179
184. Choi SH, Kang HJ, Ryu EN, Lee KP (2001) Radiat Phys Chem 60:495
185. Mardegain SB (2003) Battery Power Prod Technol 1:12
186. Boehnstedt W (1996) J Power Sources 59:45
187. Boehnstedt WJ (2004) J Power Sources 133:59
188. Lander JJ (1974) In: Proceedings of the symposium on battery separators. The Electrochemical Society, Columbus, OH, p 4
189. Prout L (1993) J Power Sources 46:117
190. Vinal GW (1945) Storage batteries. Wiley, New York
191. Butherus AD, Lindenberger WS, Vaccaro FJ (1970) Lead-acid battery: electrochemical compatibility of plastics. Bell Syst Tech J 49(7):1377–1392
192. Paik SL, Terzaghi G (1995) J Power Sources 53:283
193. Kung J (1994) J Power Sources 48:129
194. Wang LC, Harvey MK, Stein HL, Scheunemann U (1997) The role of UHMW-PE in microporous PE separators. In: Proceedings of the 12th annual battery conference on applications & advancesIEEE, New York, 14–17 Jan 1997, p 69
195. Larsen DW, Kehr CL (1996) US Patent 3,351,495
196. Boehnstedt W (2001) J Power Sources 95:234
197. Wang LC, Harvey MK, Ng JC, Scheunemann U (1998) J Power Sources 73:74
198. Endoh H (1996) J Power Sources 599:51
199. Higashi T, Endoh H (1998) J Power Sources 73:110
200. Rand DAJ, Woods R, Dell RM (1998) Batteries for electric vehicles. Research Studies Press, Taunton, UK. ISBN. ISBN 0-86380-205-0
201. McClelland DH, Devitt JL (1975) US Patent 3,862,861
202. Zguris GC (1998) J Power Sources 73:60
203. Zguris GC (1997) J Power Sources 67:307
204. Zguris GC (1996) J Power Sources 59:131
205. Pavlov D, Ruevski S, Naidenov V, Sheytanov G (2000) J Power Sources 85:164
206. Ferreira AL (1999) J Power Sources 78:41
207. Fetcenko MA, Venkatesan S, Ovshinsky S (1992) In: Proceedings of the symposium on hydrogen storage materials, batteries & electrochemistry. Electrochemical Society, Pennington, NJ, p 141
208. Wada M (1994) Polymer role in advanced battery technology. Polym Adv Tech 5:645–652
209. Ikoma M, Hoshina Y, Matsumoto L, Iwakura C (1996) J Electrochem Soc 143:1904
210. Ikoma M, Takahashi O, Tsuboi R, Matsumoto L (1993) Denki Kagaku 61:997
211. Furukawa N (1994) J Power Sources 51:45
212. Nagarajan GS, Van Zee JW (1998) J Power Sources 70:173
213. Cook JA, Lancaster IM (1998) Electrochem Soc Proc 98–15:55
214. Scimat's latest separators (2002) Batteries International, October 2002
215. Cheng S, Zhang J, Liu H, Leng Y, Yuan A, Cao C (1998) J Power Sources 74:155
216. Newman J, Tiedemann W (1975) AIChE J 21:25
217. Newman JS (1991) Electrochemical systems, 2nd edn. Prentice-Hall, Englewood Cliffs
218. Vidts PD, White RE (1997) J Electrochem Soc 144:1343
219. Ceder G, Doyle M, Arora P, Fuentes Y (2002) Computational modeling and simulation for rechargeable batteries. MRS Bull 27(8):619–623
220. Newman J, Tiedemann W (1997) J Electrochem Soc 144:3081
221. Gu H, Nguyen TV, White RE (1987) J Electrochem Soc 134:2953
222. Vidts PD, Delgado J, White RE (1995) J Electrochem Soc 142:4006
223. Doyle M, Fuller TF, Newman J (1993) J Electrochem Soc 140:1526
224. Fan D, White RE (1991) J Electrochem Soc 138:17
225. Doyle M, Newman J, Gozdz AS, Schumtz CN, Tarascon JM (1996) J Electrochem Soc 143:1890
226. Arora P, Doyle M, Gozdz AS, White RE, Newman J (2000) J Power Sources 88:219
227. Patel KK, Paulsen JM, Desilvestro J (2003) J Power Sources 122:144
228. Mao Z, White RE (1993) J Power Sources 43–44:181
229. Santhanagopalan S, Ramadass P, (Zhengming) Zhang J (2009) Analysis of internal short-circuit in a lithium ion cell. J Power Sources 194(1):550–557
230. Denton F, Howard JN, Anani AA, Fernandez JM (2001) US Patent 6,228,516
231. Takita K, kono K, Takashima T, Okamoto K (1991) US Patent 5,051,183
232. Mao H, Wainwright DS (1990) US Patent 6,074,766
233. Thomas-Alyea KE, Newman J, Chen G, Richardson TJ (2004) A concentrated solution theory model of transport in solid-polymer-electrolyte fuel cells. J Electrochem Soc 151: A509

Recreational Water Risk: Pathogens and Fecal Indicators

Alexandria B. Boehm[1], Jeffrey A. Soller[2]
[1]Environmental and Water Studies, Department of Civil and Environmental Engineering, Stanford University, Stanford, CA, USA
[2]Soller Environmental, LLC, Berkeley, CA, USA

Article Outline

Glossary
Definition of the Subject and Its Importance
Introduction
Pathogens in Coastal Waters
Assessing Risk: Epidemiology and Indicator Organisms
Assessing Risk: Quantitative Microbial Risk Assessment (QMRA) Modeling
Future Directions
Bibliography

Glossary

Microbial pollution Pathogens.

Quantitative microbial risk assessment Assessment of risk from exposure to pathogens.

Epidemiology The study of human health in response to a treatment.

Recreational waterborne illness Illness resulting from exposure to microbial pollution, includes gastroenteritis, respiratory illness, ear, nose, eye, and throat ailments, and skin rash.

Bacterial pathogens Pathogens that are bacterial.

Viral pathogens Pathogens that are viruses.

Protozoan pathogens Pathogens that are protozoa.

Definition of the Subject and Its Importance

Pathogens can enter recreational waterbodies (lakes, rivers, ocean beaches) from a number of different sources. The illness acquired by swimmers after exposure to pathogen-polluted recreational waters is termed recreational waterborne illness (RWI). Since many RWI go unreported to health care agencies, the true number of RWI each year can only be estimated; however, numbers are believed to be high – over one million in southern California alone. The risk of RWI from exposure to recreational waters can be measured using epidemiological studies and estimated using quantitative microbial risk assessments.

Introduction

Exposure to microbially polluted recreational waters can cause a variety of adverse health effects in humans including neurological infections, skin infections, earaches, eye infections, gastrointestinal illnesses, and respiratory infections [1]. Microbial pollution refers to the presence of organisms that cause illness in humans either through the production of toxins or their colonization of the human body.

It is estimated that globally, exposure to coastal waters polluted with wastewater results in an excess 120 million gastrointestinal and 50 million severe respiratory illnesses per year [2], including illnesses acquired through consumption of contaminated shellfish. In southern California, there are an estimated 1.5 million cases of gastrointestinal illnesses each year due to recreational exposure to polluted waters [3]. Moreover, there were 259 recreational water outbreaks that occurred in the USA between 1970 and 2000 [4].

Estimating the number of individuals acquiring illness through exposure to recreational waters is challenging. Typically, individuals who acquire recreational waterborne illness (RWI) do not seek medical attention because most of the illnesses tend to be mild and self-limiting. In addition, most RWIs are not reportable, so incidence levels are highly uncertain. The estimates mentioned previously were obtained using a variety of assumptions about the contamination of water and exposures. Epidemiology studies and quantitative microbial risk assessments (QMRA) are two scientifically rigorous methods that are used to estimate rates of RWI as a function of water quality.

This chapter provides a brief overview of the pathogens present in coastal waters. Readers are referred to other references for more information on these topics [5–7]. The two most common methods of assessing risk of exposure to pathogens in coastal waters, epidemiology studies, and QMRA, are also described. Example applications of these methods to assess risk of illness from exposure to pathogen-polluted coastal waters are presented. Critical research gaps are identified and summarized.

Pathogens in Coastal Waters

Pathogens present in coastal waters can be characterized into two broad groups. The first group consists of autochthonous pathogens that have an ecological niche in recreational waters. The other group is composed of allochthonous pathogens that come from human and animal wastes that have been discharged into these waterbodies.

Autochthonous pathogens include harmful algae, organisms in the genus *Vibrio*, and some protozoa such as *Naeglaria fowleri*. Harmful diatoms and dinoflagellates can cause a variety of ailments in humans from their production of toxins [8, 9]. *Vibrio vulnificus* and *V. parahaemolyticus* can infect open wounds and cause gastroenteritis if ingested from seawater. Other *Vibrio* species (such as *V. cholerae* and *V. mimicus*) can be pathogenic and cause similar ailments. *N. fowleri* is primarily found in freshwater, and when it enters the human body, can cause a rare but serious brain infection. While these organisms are extremely important from a human health perspective, this chapter will not discuss risks associated with autochthonous pathogens.

The focus of this chapter is allochthonous pathogens (Table 1) including the eight pathogens that cause a large proportion (>95%) of all non-foodborne illnesses in the USA [10]: the viruses norovirus, rotavirus, and adenovirus; the bacteria *Campylobacter*, *Salmonella* and pathogenic *Escherichia coli*; and the protozoans *Cryptosporidium* spp. and *Giardia lamblia*. Other pathogens that are important etiologies of RWI include enteroviruses, *Staphylococcus aureus,* and methicillin-resistant *S. aureus* (MRSA). All these pathogens, with the exception of *S. aureus* and MRSA, cause gastrointestinal illness. *Staphylococcus* causes skin infections.

Recreational Water Risk: Pathogens and Fecal Indicators. Table 1 Allothchonous human pathogens detected in coastal waters

	Concentration/Occurrence	Reference
Viruses		
Enteroviruses	Present in 9 of 72 1-liter samples using RT-PCR at Avalon Beach, CA[a]	[81]
Adenoviruses	Present in 15 of 30 250-liter samples using PCR at Silver Beach, MI[a]	[63]
Hepatitis A	105–30,771 viral particles/l using Q-RTPCR at Imperial Beach, CA[a]	[82]
Norovirus	2 of 19 samples in 110-liters using RT-PCR at Key West sites (FL)[a]	[83]
Rotavirus (reovirus)	2 of 19 sites with 2–5 MPN/L at Italian coastline	[84]
Bacteria		
Campylobacter	Detected in 25 of 192 100 – 1,000 mL Spanish marine recreational water samples using culture-based methods	[85]
Salmonella	Detected in 70–100% of samples from a lagoon in Brazil using culture-based methods, volume assayed not reported	[86]
Staphylococcus	60–70% of approx. 100 mL seawater samples from Doheny and Avalon Beach, CA using culture-based methods	[13]
Pathogenic *Escherichia coli*	2 of 377 *E. coli* isolates from North Carolina and Southern California coastal waters using combined culture and PCR methods	[87]
Shigella	100% of algal mat samples from Lake Michigan near Burns Ditch by PCR	[18]
Protozoa		
Cryptosporidium	13.7 ± 1.7 oocysts/L on weekends at Chesapeake Bay beach, MD	[88]
Giardia	9.1 ± 1.1 cysts/L on weekends at Chesapeake Bay beach, MD	[88]

[a] Volumes reported do not account for the fact that a fraction of water sample was used during polymerase chain reaction (PCR), reverse-transcriptase (RT-)PCR, or quantitative (Q)PCR.

Adenoviruses can cause gastrointestinal illness, as well as eye and respiratory infections.

The detection of pathogens in environmental matrices is methodologically challenging. Allochthonous pathogens are rare microbes in the environment. In seawater there are on the order of one million autochthonous bacteria and 10 million autochthonous viruses in a milliliter. Whereas allochthonous pathogens may be present at levels of 1 per 10 liters l or lower. Thus, enumerating the allochthonous pathogens is particularly difficult because the presence of all the other organisms in the sample can interfere with the detection of the rare target. Because the field of pathogen detection is still evolving and because allochthonous pathogens densities are typically low, there are limited data on pathogen concentrations and occurrence in recreational waters. In many cases, authors only provide data on the presence or absence of pathogens in recreational waters and do not provide concentrations. Some example concentrations are provided in Table 1.

Treated and untreated wastewater, human and other animal feces, stormwater and urban runoff, and agricultural runoff can all contain microbes that are pathogenic to humans [11, 12]. When discharged to coastal waters, concentrations of pathogens may be high and can pose high levels of human health risks.

Pathogens are also found in environmental reservoirs that in some cases may serve as a source of pathogens to recreational waters. Beach sands can harbor *Campylobacter, Salmonella*, and *S. aureus* [13–16]. Aquatic sediments can accumulate bacteria, protozoa, and viruses [17]. Marine and lacustrine kelp species may also harbor bacterial pathogens including pathogenic *E. coli*, *Salmonella*, and *Campylobacter* [18, 19].

Assessing Risk: Epidemiology and Indicator Organisms

The monitoring of recreational waters for all RWI pathogens to assess the safety of swimming is not scientifically or economically feasible [20]. Microbial indicator organisms have been used for centuries as indicators of the presence of human pathogens. Internationally, many countries use fecal and total coliforms as a basis for their recreational water quality criteria, standards, or guidelines [21]. Other countries rely on measurements of enterococci (or fecal streptococci), *E. coli*, or both for their recreational waters, most based on guidelines provided by World Health Organization [22] and/or the US Environmental Protection Agency (EPA). These organisms were chosen as indicators because their concentrations are high in human wastewater and feces, they are relatively simple to measure, and their presence in coastal waters is correlated to adverse health outcomes in swimmers through epidemiology studies conducted in wastewater-impacted waters [23–26]. The epidemiology studies that correlate indicator concentration to adverse health outcomes are key to the use of indicators to assess risk.

Epidemiology studies evaluate illness resulting from exposure to a particular contaminant or activity. When applied to RWI, epidemiology studies evaluate the illness rates in swimmers versus non-swimmers, and characterize illness rates as a function of indicator organism concentration. The studies involve the collection of health and behavior data contemporaneously with concentrations of indicator organisms. RWI epidemiology studies are either case-control randomized trial or prospective cohort designs. In case-control studies, swimmer and non-swimmer activities are prescribed by randomization at the onset of the study. In these studies, exposures are well controlled. Subject recruitment is done in advance of the study. In prospective cohort studies, subjects are recruited at the study site and are enrolled when they arrive at the shoreline. Exposures are self-prescribed by subjects and behavioral data on exposure is collected using self-reports at the end of the day. In the prospective cohort design, there is less control over exposure, but the exposures are more realistic as they are not prescribed by the study design. Additional types of studies that have been employed to study RWI include cross-sectional studies and event studies. The former is similar in design to a cohort study; the latter takes advantage of a sporting event, for example, for data collection.

Since the 1950s, numerous epidemiological studies have been conducted throughout the world to evaluate the association between recreational water quality and RWI (including GI symptoms; eye infections; skin complaints; ear, nose, and throat infections; and respiratory illness) [23–26]. Most of these studies investigated wastewater effluent-impacted marine and estuarine waters alone or in combination with freshwater. Several investigated freshwater recreational environments or non-wastewater effluent-impacted waters. These studies indicate that the rates of some adverse health outcomes are higher in swimmers compared with non-swimmers [23].

Taken as a whole, the weight of evidence from these studies indicates that fecal indicator bacteria (fecal streptococcus/*Enterococcus*, in particular) are able to predict GI and in some cases, respiratory illnesses from exposure to recreational waters [23, 25, 26]. This broad base of information stems from studies conducted throughout much of the developed world (Table 2, Adapted from [27]).

Several meta-analyses and/or systematic reviews have summarized the available recreational water epidemiology studies [23, 25, 26]. Pruss et al. [23] conducted a systematic review to initiate development of new WHO guidelines for recreational use of the water environment. The comprehensive review of 22 published studies on sewage pollution of recreational water and health outcomes concluded that the epidemiological basis had been laid to develop WHO guidelines on fecal pollution based on a causal association between GI illness symptoms and increased concentrations of bacterial indicators (i.e., enterococci for marine, enterococci and *E. coli* for fresh) in recreational waters.

Zmirou et al. [26] examined 18 published studies to provide a scientific basis for establishing new standards for the microbial quality of marine and fresh recreational waters to replace the 30 year-old European Union bathing water quality guidelines [28]. The researchers provided four major results: (1) increased concentrations of fecal coliforms or *E. coli* and enterococci in both fresh and marine recreational waters are associated with increased risks of acute GI illness, with enterococci eliciting four to eight times greater excess risks than fecal coliforms or *E. coli* at the same concentrations; (2) GI illness risks associated with enterococci occur at lower indicator concentrations in marine versus fresh recreational waters; (3) increased concentrations of total coliforms have little or no association with GI illness risk; and (4) no evidence exists of a concentration threshold of indicator microorganisms below which there would be no GI illness risk to bathers.

Wade et al. [25] conducted a systematic review and meta-analysis of 27 published studies to evaluate the evidence linking specific microbial indicators of recreational water quality to specific health outcomes under non-outbreak (endemic) conditions. The study was conducted at the request of the United States National Academy of Sciences. Secondary goals included identifying and describing critical study design issues and evaluating the potential for health effects at or below the current regulatory criteria [29]. They concluded that (1) enterococci and, to a lesser extent, *E. coli* are adequate indicators (predictors) of GI illness in marine recreational waters, but fecal coliforms are not; (2) the risk of GI illness is considerably lower in studies with enterococci and *E. coli* densities below those established by EPA [29], thus providing support for their regulatory use; (3) *E. coli* is a more reliable and consistent predictor of GI illness than enterococci or other indicators in fresh recreational waters; and (4) studies that used a non-swimming control group and that focused on children found elevated GI illness risks.

Based on these meta-analyses, the weight of evidence indicates that there is a relationship between levels of specific indicator bacteria and RWI in coastal waters impacted by wastewater. However, as discussed earlier there are many sources of indicator bacteria to coastal waters, and many of these sources contain different pathogens with diverse health risks. Along coastlines with good sewage infrastructure and regulated anthropogenic discharges, wastewater is unlikely to be contributing substantial amounts of indicator organisms to the swimming areas on a regular basis [27]. Important sources of indicator organisms and pathogens are probably nonpoint in nature, emanating from soils, animal feces, urban runoff, stormwater runoff, or agricultural runoff.

There are only a few epidemiology studies that examine the link between RWI and fecal indicator organisms in recreational waters polluted with sources other than wastewater. Review of these studies suggests the relationship between indicator concentration and RWI risks are equivocal. On one hand, Colford et al. [30] found that the incidence of swimmer illness was greater than the incidence of non-swimmer illness, but swimmer illness was not associated with any of the bacterial indicator organisms at a marine beach where bacterial contamination was not attributable to

Recreational Water Risk: Pathogens and Fecal Indicators. Table 2 Recreational water epidemiology studies included in reviews by Prüss [23], Wade et al. [25], and Zmirou et al. [26]. Location refers to geographic location of study. Water Type refers to whether the study was conducted at a marine or fresh water. Study design denotes whether to study was a cohort, randomized trial, cross section, or event study

Reference	Location	Water type	Study design	Review article
Alexander et al. [89]	UK	Marine	Cohort	Wade, Zmirou
Bandaranayake [90]	New Zealand	Marine	Cohort	Prüss
Brown et al. [49]	UK	Marine	Cohort	Zmirou
Cabelli [38]	USA	Marine	Cohort	Wade, Prüss, Zmirou
Cabelli [44]	Egypt	Marine	Cohort	Wade, Prüss
Calderon et al. [31]	USA	Fresh	Cohort	Wade
Cheung et al. [36]	Hong Kong	Marine	Cohort	Wade, Prüss, Zmirou
Corbett et al. [52]	Australia	Marine	Cohort	Wade, Prüss, Zmirou
Dufour [39]	USA	Fresh	Cohort	Wade, Prüss, Zmirou
Fattal et al. [42]	Israel	Marine	Cohort	Wade, Prüss, Zmirou
Ferley et al. [47]	France	Fresh	Cohort	Wade, Prüss, Zmirou
Fewtrell et al. [91]	UK	Fresh	Event	Wade, Zmirou
Fewtrell et al. [92]	UK	Marine	Cohort	Wade, Zmirou
Kay et al. [71]	UK	Marine	Randomized trial	Wade
Fleisher et al. [93]	UK	Marine	Randomized trial	Prüss
Foulon et al. [48]	France	Marine	Cross-sectional	Wade
Haile et al. [34, 94]	USA	Marine	Cohort	Wade, Prüss, Zmirou
Kay et al. [71]	UK	Marine	Randomized trial	Wade, Prüss, Zmirou
Kueh et al. [95]	Hong Kong	Marine	Cohort	Wade, Prüss
Lee et al. [96]	UK	Fresh	Event	Wade
Lightfoot [97]	Canada	Fresh	Cohort	Wade, Prüss
Marino et al. [98]	Spain	Marine	Cohort	Wade
McBride et al. [33]	New Zealand	Marine	Cohort	Wade
Medema et al. [99]	The Netherlands	Fresh	Event	Wade
CSIR [100]	South Africa	Marine	Cohort	Prüss
Mujeriego et al. [46]	Spain	Marine	Cohort	Prüss
Philipp et al. [101]	UK	Marine	Event	Wade, Zmirou
Pike [102]	UK	Marine	Cohort	Wade, Prüss, Zmirou
Prieto et al. [103]	Spain	Marine	Cohort	Wade
Seyfried et al. [40]	Canada	Fresh	Cohort	Wade, Prüss, Zmirou
Stevenson [104]	USA	Fresh	Cohort	Wade, Prüss, Zmirou
UNEP/WHO [105]	Israel	Marine	Cohort	Prüss
UNEP/WHO [106]	Spain	Marine	Cohort	Prüss

Recreational Water Risk: Pathogens and Fecal Indicators. Table 2 (Continued)

Reference	Location	Water type	Study design	Review article
Van Asperen et al. [107]	The Netherlands	Fresh	Event	Wade, Zmirou
Van Dijk et al. [108]	UK	Marine	Cohort	Prüss
Von Schirnding et al. [109]	South Africa	Marine	Cohort	Wade, Zmirou

wastewater discharges. Similarly, Calderon et al. [31] found no statistically significant association between swimmers' illness risk and indicator concentrations in a freshwater pond where agricultural runoff was the source of contamination. McBride [32] suggested that if more swimmers had been included in the Calderon et al. [31] study, achieving statistically significant results would have been possible, however. At a marine bathing study in New Zealand, McBride et al. [33] indicated that RWI risks posed by animal versus human fecal material were not substantially different; however, the study's limited range of indicator organisms' concentrations precluded the development of a detailed statistical model of health risks versus indicator density.

In the first study to be conducted in waters directly impacted by urban runoff, Haile et al. [34] reported associations between swimmer health and indicator densities. However, this nonpoint runoff source was known to contain human sources of fecal contamination, based on the presence of human enteric viruses. Dwight and colleagues [35] found that surfers exposed to Southern California urban runoff had higher illness rates than surfers exposed to Northern California rural runoff. The results from a Hong Kong marine water study [36] and a German freshwater study [37] are more difficult to interpret regarding risks from human versus nonhuman sources because in both studies, the analyses combined the results from sites with different predominant contamination sources.

An additional meta-analysis examined the differential risks associated with exposure to human and nonhuman animal fecal material [24]. Illness risk associated with bathing in water polluted primarily with human fecal material was reviewed based on studies from the USA [38, 39], Canada [40, 41], Israel [42, 43], Egypt [44, 45], Spain [46], France [47, 48]; the UK [49, 50], Hong Kong [51], and Australia [52, 53]. Most of these studies showed a positive correlation between GI illness and fecal indicator density; there was little equivalent evidence from waters polluted primarily with animal feces. The only study specifically designed to address swimming-associated illness in animal-impacted waters was that of Calderon et al. [31] who found no statistically significant association between GI illness and fecal indicator bacteria densities. Based on this observation, Sinton and colleagues [24] concluded that reliable epidemiological evidence was lacking, and that other sources of information were needed to identify and apportion human and nonhuman fecal inputs to natural waters.

Given these studies, there are not sufficient epidemiological data to conclude that concentration of indicator organisms in coastal waters not impacted by wastewater are predictive of health risks in all cases. More epidemiology data may help address the lack of information. However, fecal contamination in coastal waters not impacted by wastewater is likely to be highly variable and emanate from a complex mix of sources. Other approaches may be more useful for addressing the relationship between indicators and health risk along these shorelines. Despite the lack of epidemiological evidence for a relationship between indicator organism concentration and health risk, it is well established in the outbreak literature that water contaminated with animal feces can cause illness in humans [4].

Assessing Risk: Quantitative Microbial Risk Assessment (QMRA) Modeling

QMRA is a health risk modeling approach that translates microbial exposures into infection or illness risk estimates. For RWI, the dose received by individuals is derived from estimates of the volume of water ingested during an exposure event and concentration of pathogen(s) in that volume of water. Once a dose (number of pathogens per exposure) is determined then a risk of infection or illness is derived from applying published dose-response models for specific pathogens [54], which are derived from human feeding trials or outbreak data [55, 56].

Estimations of key model parameters (such as ingested volume of water and pathogen concentration) are generally described as probability density functions (PDFs) (i.e., distributions to account for the stochastic nature of the parameter) to help account for inherent variability as well as various methods and model uncertainties. The characterized risk is best described as a distribution as well, to capture variability and uncertainty as best as possible. This characterized risk can then be compared to "tolerable risk" or a site benchmark for the recreational water (e.g., 8 or 19 gastrointestinal illnesses per 1,000 bathers as used in the 1986 EPA criteria for freshwater and marine recreation, respectively).

An initial screening-level risk assessment for a site may start by only using point estimates to describe model parameters (e.g., WHO, 2004). Then, to reduce uncertainties in the risk estimate, more complexity is built into a QMRA model using PDFs to better represent a specific site, in what is called a "static" model. An alternative approach known as "dynamic QMRA" takes into account secondary infections to the broader community, as well as addressing the susceptibility versus nonsusceptibility of individuals to infection and illness [57–59].

QMRA can be useful for a number of purposes. First, it can be used to look at hypothetical risks under different scenarios of pathogen sources and/or recreational activities and exposure routes [60, 61]. This can provide a human health interpretation of environmental pathogen concentrations, or it can provide guidance for decision-making with respect to alternative management options. Second, QMRA can be used to augment the understanding of recreational water epidemiology studies [62]. A number of studies have evaluated pathogen risks in recreational waters using QMRA [60, 61, 63–70]. A handful of these are reviewed below.

Diallo et al. [69] examined the risk associated with recreational exposure to canals in Thailand containing *Giardia*, *Cryptosporidum*, and diarrhegenic *E. coli*. They found the predicted risk of illness from the protozoa was two orders of magnitude higher than the actual protozoan infection rate in the region of Thailand, while the rate of gastrointestinal illness predicted from diarrhegenic *E. coli* matched actual observed rates of the disease in the region. The authors suggest that the illness rates predicted for the protozoa are much higher due to immunity in the community, which was not considered in the QMRA modeling. Another possibility is massive under-reporting of protozoa illness rates, which is a documented issue for diarrheal diseases.

Ashbolt and Bruno [65] used QMRA in conjunction with the published relationship between enterococci concentrations and probability of gastrointestinal illness and acute respiratory illness observed during a UK beach epidemiology study [71]. The authors showed that by assuming that the etiology of illnesses was viral, a fixed ratio between enterococci and viruses of 1–175, a volume of water ingested of 50 mL, and 50% illness rate of those infected with virus, they were able to model the observed illness rates (at exposures greater than 60 CFU/100 mL enterococci) assuming viruses had an exponential dose-response curve. Further, they were able to model the observed acute respiratory illness using the dose-response curve of adenovirus.

Gerba et al. [70] estimated risk of exposure to rotavirus in bathing waters using the few previously reported data available at the time on rotavirus concentrations in bathing waters. They found the risk of illness to be 1/10–1/100 with a one time bathing event. These authors chose to use rotavirus as a model pathogen because it is one of the most infectious viruses for which a dose-response model is available. A major limitation to their study is the lack of data on concentrations of rotavirus in bathing waters. Surface water concentrations were estimated to be between 0.24/L and 29/L. Ingested volumes used were 100 mL for recreational exposure.

Schoen and Ashbolt [67] explored the relative risk associated with exposure to seagull feces, poorly treated sewage, and a mixture of both sources in seawater with enterococci at levels of 35 CFU/100 mL (a USEPA standard). Authors assumed that seagull feces contained *Campylobacter* and *Salmonella*, while sewage contained norovirus, *G. intestinalis*, *Cryptosporidium* spp., and *Salmonella*. A distribution of ratios of enterococci to pathogen concentrations were considered to reflect the uncertainty in this parameter. The authors showed that when enterococci came exclusively from seagull feces, the risk of illness is less than the benchmark of 0.01 on which US standards are based.

Soller et al. [61] examined the risk of viral gastroenteritis associated with recreational and non-recreational use of a river downstream of a wastewater treatment plant discharge. Two wastewater treatment scenarios were compared with the goal of evaluating the public health benefit of increased treatment of the effluent. The authors employed a health protective approach by assuming that the etiology of illness would be a virus with clinical features identical to those of rotavirus. They also assumed that removal of the virus in treatment facilities would mirror that of coliphage and that coliphage and the virus occurred in a ratio that varied from 0.001 to 1. Exposures were informed by a hydrologic model of the area and observations of swimmer behavior. Unlike the other models discussed above, this model incorporated secondary transmission by allowing illness to be passed from person to person.

Steyn et al. [68] compared the risks in recreational surfaces waters in South Africa expected from measured *E. coli* concentrations (applying epidemiology study-derived dose-response curves) and measured *Salmonella* concentrations (applying the QMRA method and the *Salmonella* dose-response curve assuming ingestion of 100 mL of water for exposure). The researchers found that risks derived from the *Salmonella* QMRA model were higher than those derived from the *E. coli* concentrations. Their results suggest using *E. coli* to assess risk from exposure to *Salmonella* may be inadequate.

Wong et al. [63] estimated risk from exposure to adenoviruses during recreation contact with water at Great Lakes beaches impacted by point sources of treated wastewater. The authors measured adenoviruses using an MPN culture-dependent assay at their study site, assumed an adenovirus dose-response model, and an ingestion rate of 100 mL of water. The authors found that 0.24–2.4 illnesses per 1,000 swimmers were likely to have occurred from adenoviruses, a range of frequencies below the EPA guideline of 8 illnesses per 1,000 swimmers. Although an epidemiology study was done at the same time as their study, the epidemiology data were not available for comparison of actual illness rates.

Soller et al. [62] used QMRA to understand more fully the reported epidemiologic results from studies conducted on the Great Lakes in the US during 2003 and 2004 by identifying pathogens that could have caused the observed illnesses in those studies. The reference pathogens used for this analysis were *Norovirus*, rotavirus, adenovirus, *Cryptosporidium* spp., *G. lamblia, Campylobacter jejuni*, *Salmonella enterica, and E. coli* O157:H7. Two QMRA-based approaches were used to estimate the pathogen combinations that would be consistent with observed illness rates: in the first, swimming-associated gastrointestinal (GI) illnesses were assumed to occur in the same proportion as known illnesses in the US due to all non-foodborne sources, and in the second, pathogens were assumed to occur in the recreational waters in the same proportion as they occur in disinfected secondary effluent. The results indicated that human enteric viruses and in particular, norovirus could have caused the vast majority of the observed swimming-associated GI illnesses during the 2003/2004 water epidemiology studies [72, 73]. Evaluation of the time to onset of illness strongly supports the principal finding, and sensitivity analyses support the overall trends of the analyses even given their substantial uncertainties. These results are notable because little is known about the specific microbial agents that are responsible for the observed illnesses in swimmers. While several studies have attempted to collect biological specimens (blood or stool) as part of epidemiologic research at beach sites, these efforts have to date been largely unsuccessful in identifying the agents responsible for the observed increase in GI symptoms among swimmers [50].

Soller et al. [74] conducted a QMRA investigation to determine whether estimated risks following exposure to recreational waters impacted by gull, poultry, pig, or cattle fecal contamination are substantially different from those associated with waters impacted by human sources such as treated wastewater. Previously published QMRA methods were employed and extended to meet these objectives [67]. Health outcomes used in the analyses were infection via reference pathogens by water ingestion during recreation and subsequent GI illness. Illness risks from the reference pathogens were calculated for exposure to contaminated recreational water with fecal indicator bacterial densities at the U.S. regulatory limits: 35 CFU/100 mL enterococci and 126 CFU/100 mL *E. coli*. The probabilities of GI illness from reference pathogens were calculated using dose-response relationships from the literature and Monte Carlo simulations. The primary findings from the analysis were that: (1) GI illness risks associated with human exposure to recreational waters directly impacted by fresh cattle feces are not substantially different from those impacted by human sources; and (2) the risks associated with human exposure to recreational waters impacted by fresh gull, poultry, or pig feces are substantially lower than those impacted by human sources.

Several observations may be drawn from the QMRA studies summarized above. First, the assembled studies focused on a small subset of the pathogens potentially important in waterborne exposure during recreation. The pathogen analyzed most frequently is rotavirus, primarily due to its high infectivity and the availability of dose-response data. It is likely that future QMRAs will also focus on norovirus [62, 67, 74] now that a published dose-response relationship is available [75].

Second, modeling variability in pathogen source density appears to be hampered by scarcity of both data and analysis techniques. Two common methods for accounting for source variability among studies are (1) use of empirical distributions for pathogen density based on relatively short time series, and (2) assumption of log-normal distribution of pathogen densities. There are advantages and drawbacks to both of these approaches. Drawbacks to the use of empirical distributions are inconsistency in sampling strategies used to develop databases, frequent non-detects, constraint of pathogen densities to those observed in a limited number of samples, and lack of availability of pathogen concentration data. Use of distributions to describe pathogen density in sources overcomes some of the constraints associated with use of empirical distributions, but choosing a distributional family can be problematic. Among the studies reviewed, many studies employed point estimates for pathogen density. Among studies using distributions to describe pathogen variability, the following distributions were employed: normal, triangular, log-normal, negative binomial, uniform, and Poisson.

Third, most of the studies reviewed do not account for variability and uncertainty in dose-response model parameters. As with variability in exposure, this observation likely indicates that high quality and diverse dose-response model data are not available. The use of a small number of dose-response models may indicate that some QMRA modelers chose the pathogens to model based on the availability of dose-response models. Lack of dose-response models for many pathogens of concern and for differing routes of exposure (i.e., cutaneous exposure to *S. aureus*) is a major data gap. The need for dose-response models corresponding to different exposure routes (i.e., ingestion, inhalation, etc.) arises from the ability of some waterborne pathogens such as adenovirus to initiate infection via multiple routes.

Finally, secondary transmission and immunity are often neglected in risk estimation. Studies that have included secondary transmission and waterborne illnesses [57, 58, 60, 76–78] have demonstrated that consideration of secondary transmission and immunity can influence overall risk associated with exposure to pathogens significantly and in unintuitive ways.

Future Directions

Health data collected from recreational swimmers confirm measurable health effects from exposure to contaminated coastal waters. To adequately protect the health of swimmers and others who recreate in coastal waters or consume fish and shellfish from coastal waters, it is essential to understand the microbial hazards present and the risks they represent to human health. While there are data available on microbial hazards and risks, many aspects remain poorly understood and characterized.

Research is needed to further characterize microbial hazards in recreational waters. Pathogen detection techniques that allow detection of infectious pathogens rapidly in environmental waters are needed. There are a number of pathogen and indicator detection technologies that are in development or have been recently developed [79, 80]. However, many of these have not been applied to natural waters, or are just in the "proof of concept" stage. Work is needed to transit new detection technologies that work at the bench-scale to the field scale – to detect pathogens in environmental waters. Once this is accomplished, they should be applied to a wide array of waterbodies to fully understand pathogen and indicator occurrence, concentrations, fate, and transport.

A better understanding of the risks of exposure to pathogens from different sources is needed. Some QMRA studies have addressed this issue, but further research is needed to ground truth the QMRA results with epidemiology data, or data on infection rates and etiologies, as determined through analysis of bodily fluid from individuals with RWI. Gastroenteritis is the most well-studied RWI; more work is needed to understand the importance of other RWI including skin infections and respiratory ailments. More dose-response data for a wider array of pathogens are needed to provide more refined estimates of risk using QMRA. The importance of secondary infections and immunity for RWI should also be further characterized.

Finally, better surveillance systems are needed for RWI so that prediction systems can be developed. As global climate changes in the coming decades, the scientific community needs to be able to anticipate how this might change the burden of RWI. Microbial pollution of coastal waters is expected to change as temperatures, runoff frequency and volumes, and rainfall pattern change. A thorough understanding of the occurrence of RWI, and the distribution, fate, and transport of waterborne pathogens will enable us to better anticipate the effects of climate change.

Bibliography

1. CDC http://www.cdc.gov/healthywater/swimming/rwi.html (5 January)
2. Shuval H (2003) Estimating the global burden of thalassogenic diseases: human infectious diseases caused by wastewater pollution in the environment. J Water Health 1:53–64
3. Given S, Pendleton LH, Boehm AB (2006) Regional public health cost estimates of contaminated coastal waters: a case study of gastroenteritis at Southern California beaches. Environ Sci Technol 40:4851–4858
4. Craun GF (2003) Causes of waterborne outbreaks in the United States. Wat Sci Tech 24:17
5. Santo-Domingo JW, Hansel J (2008) Waterborne diseases and microbial water quality monitoring for recreational water bodies using regulatory methods. In: Walsh PJ, Smith SL, Fleming LE, Solo-Gabriele HM, Gerwick WH (eds) Oceans and human health. Academic, San Diego
6. Griffin DW, Donaldson KA, Paul JH, Rose JB (2003) Pathogenic human viruses in coastal waters. Clin Microbiol Rev 16(1):129–143
7. Maier RM, Pepper IL, Gerba CP (2000) Environmental microbiology. Academic, San Diego
8. Boehm AB, Bischel HN (2010) Oceans and human health. In: Nriagu J (ed) Encyclopedia of environmental health. Elsevier, New York, NY
9. Walsh PJ, Smith SL, Fleming LE, Solo-Gabriele HM, Gerwick WH (2008) Oceans and human health. Academic, San Diego
10. Mead PS, Slutsker L, Dietz V, McCaig LF, Bresee JS, Shapiro C, Griffin PM, Tauxe RV (1999) Food related illness and death in the United States. Emerg Infect Dis 5(5):607–625
11. Olivieri AW, Boehm AB, Sommers CA, Soller JA, Eisenberg JNS, Danielson R (2007) Development of a protocol for risk assessment of separate stormwater system microorganisms; Water Environment Research Foundation. Project 03-SW-2, Final Project Report
12. U.S. EPA (2009) Review of published studies to characterize relative risks from different sources of fecal contamination in recreational water, EPA 822-R-09-001. In Lett Appl Microbiol, Office of Science and Technology, Washington
13. Goodwin KD, Pobuda M (2009) Performance of CHROMagar™ Staph aureus and CHROMagar™ MRSA for detection of Staphylococcus aureus in seawater and beach sand – Comparison of culture, agglutination, and molecular analyses. Water Res 43(19):4802–4811
14. Bolton F, Surman SB, Martin K, Wareing DR, Humphrey TJ (1999) Presence of Campylobacter and Salmonella in sand from bathing beaches. Epidemiol Infect 122:7–13
15. Elmanama AA, Fahd MI, Afifi S, Abdallah S, Bahr S (2004) Microbiological beach sand quality in Gaza Strip in comparison to seawater quality. Environ Res 99:1–10
16. Obiri-Danso K, Jones K (2000) Intertidal sediments as reservoirs for hippurate negative campylobacters, salmonellae and faecal indicators in three EU recognized bathing waters in North West England. Water Res 34:519–527
17. LaBelle RL, Gerba CP, Goyal SM, Melnick JL, Cech I, Bogdan GF (1980) Relationships between environmental factors, bacterial indicators, and the occurrence of enteric viruses in estuarine sediments. Appl Environ Microbiol 39(3):588–596
18. Ishii S, Yan T, Shivley DA, Byappanahalli MN, Whitman RL, Sadowsky MJ (2006) Cladophora (Chlorophyta) spp. harbor human bacterial pathogens in nearshore water of Lake Michigan. Appl Environ Microbiol 72:4545–4553
19. Byappanahalli MN, Sawdey R, Ishii S, Shively DA, Ferguson JA, Whitman RL, Sadowsky MJ (2009) Seasonal stability of Cladophora-associated Salmonella in Lake Michigan watersheds. Water Res 43:806–814
20. U.S. EPA (2009) Review of published studies to characterize relative risks from different sources of fecal contamination in recreational water, EPA 822-R-09-001. In Office of Science and Technology, Washington
21. WHO (1999) Health-based monitoring of recreational waters: the feasibility of a new approach (The 'Annapolis Protocol'); WHO/SDE/WSH/99.1, World Health Organization, 1999

22. WHO (2003) Guidelines for safe recreational-water environments, coastal and fresh-waters, vol 1. World Health Organization, Geneva
23. Prüss A (1998) Review of epidemiological studies on health effects from exposure to recreational water. Int J Epidemiol 27(1):1–9
24. Sinton LW, Finlay RK, Hnnah DJ (1998) Distinguishing human from animal faecal contamination in water: a review. NZ J Mar Freshw Res 32:323–348
25. Wade TJ, Pai N, Eisenberg J, Colford JM (2003) Do US EPA water quality guidelines for recreational waters prevent gastrointestinal illness? A systematic review and meta-analysis. Environ Health Perspect 111(8):1102–1109
26. Zmirou D, Pena L, Ledrans M, Letertre A (2003) Risks associated with the microbiological quality of bodies of fresh and marine water used for recreational purposes: summary estimates based on published epidemiological studies. Arch Environ Health 58(11):703–711
27. Boehm AB, Ashbolt NJ, Colford JM, Dunbar J, Fleming LE, Gold MA, Hansel JA, Hunter PR, Ichida AM, McGee CD, Soller JA, Weisberg SB (2009) A sea change ahead for recreational water quality criteria. J Water Health 7:9–20
28. EEC (1976) Bathing water quality: Directive 76/160/EEC, 1–6
29. U.S. EPA (1986) Ambient water quality criteria for bacteria, EPA440/5-84-002, Office of Water, Washington
30. Colford JM Jr, Wade TJ, Schiff KC, Wright CC, Griffith JF, Sandhu SK, Burns S, Sobsey M, Lovelace G, Weisberg SB (2007) Water quality indicators and the risk of illness at beaches with nonpoint sources of fecal contamination. Epidemiology 18(1):27–35
31. Calderon R, Mood E, Dufour A (1991) Health effects of swimmers and nonpoint sources of contaminated water. Int J Environ Health Res 1:21–31
32. McBride GB, Calderon RL, Mood EW, Dufour AP (1991) Health effects of swimmers and non-point sources of contaminated water. Int J Environ Health Res 1:21–31
33. McBride GB, Salmond CE, Bandaranayake DR, Turner SJ, Lewis GD, Till DG (1998) Health effects of marine bathing in New Zealand. Int J Environ Health Res 8(3):173–189
34. Haile RW, Witte JS, Gold M, Cressey R, McGee C, Millikan RC, Glasser A, Harawa N, Ervin C, Harmon P, Harper J, Dermand J, Alamillo J, Barrett K, Nides M, Wang GY (1999) The health effects of swimming in ocean water contaminated by storm drain runoff. Epidemiology 10(4):355–363
35. Dwight RH, Baker DB, Semenza JC, Olson BH (2004) Health effects associated with recreational coastal water use: urban versus rural California. Am J Public Health 94(4):565–567
36. Cheung WH, Chang KC, Hung RP, Kleevens JW (1990) Health effects of beach water pollution in Hong Kong. Epidemiol Infect 105(1):139–162
37. Wiedenmann A, Kruger P, Dietz K, Lopez-Pila JM, Szewzyk R, Botzenhart K (2006) A randomized controlled trial assessing infectious disease risks from bathing in fresh recreational waters in relation to the concentration of Escherichia coli, intestinal enterococci, Clostridium perfringens, and somatic coliphages. Environ Health Perspect 114(2):228–236
38. Cabelli VJ (1983) Health effects criteria for marine recreational waters, EPA-600/1-80-031. USEPA, Cincinnati
39. Dufour A (1984) Health effects criteria for fresh recreational waters; EPA-600 1-84-004. US EPA, Cincinnati
40. Seyfried PL, Tobin RS, Brown NE, Ness PF (1985) A prospective study of swimming-related illness. I. Swimming-associated health risk. Am J Public Health 75(9):1068–1070
41. Seyfried PL, Tobin RS, Brown NE, Ness PF (1985) A prospective study of swimming-related illness. II. Morbidity and the microbiological quality of water. Am J Public Health 75(9):1071–1075
42. Fattal B, Peleg-Olevsky E, Yoshpe-Purer Y, Shuval H (1986) Association between morbidity among bathers and microbial quality of seawater. Water Sci Technol 18(11):59–69
43. Fattal DB, Peleg-Olevsky E, Agursky T, Shuval PHI (1987) The association between seawater pollution as measured by bacterial indicators and morbidity among bathers at Mediterranean bathing beaches of Israel. Chemosphere 16(2/3):565–570
44. Cabelli V (1983) Public health and water quality significance of viral diseases transmitted by drinking water and recreational water. Water Sci Technol 15:1–15
45. El-Sharkawi HF (1979) The relation between the state of pollution in Alexandria swimming beaches and the occurrence of typhoid among bathers. Bull High Inst Public Health 9:337–351
46. Mujeriego R, Bravo J, Feliu M (1982) Recreation in coastal waters: public health implications. Vier Journee Etudes Pollutions, 585–594
47. Ferley JP, Zmirou D, Balducci F, Baleux B, Fera P, Larbaigt G, Jacq E, Moissonnier B, Blineau A, Boudot J (1989) Epidemiological significance of microbiological pollution criteria for river recreational waters. Int J Epidemiol 18(1):198–205
48. Foulon G, Maurin J, Quoi N, Martin-Boyer G (1983) Relationship between the microbial quality of water and health effects. Revue des Sciences de L’Eau 2:127–143
49. Brown JM, Campbell EA, Rickards AD, Wheeler D (1987) Sewage pollution of bathing water. Lancet 2(8569): 1208–1209
50. Jones F, Kay D, Stanwell-Smith R, Wyer M (1991) Results of the first pilot-scale controlled cohort epidemiological investigation into the possible health effects of bathing in seawater at Langland Bay, Swansea. Water Environ J 5(1):91–98
51. Holmes P (1989) Research into health risks at bathing beaches in Hong Kong. J Inst Water Environ Manage 3:488–495
52. Corbett SJ, Rubin GL, Curry GK, Kleinbaum DG (1993) The health effects of swimming at Sydney beaches. The Sydney beach users study advisory group. Am J Public Health 83(12):1701–1706
53. Harrington JF, Wilcox DN, Giles PS, Ashbolt NJ, Evans JC, Kirton HC (1993) The health of Sydney surfers: an epidemiological study. Water Sci Technol 27:175–181

54. Haas CN (1983) Estimation of risk due to low doses of microorganisms: a comparison of alternative methodologies. Am J Epidemiol 55:573–582
55. Haas CN, Rose JB, Gerba CP (1999) Quantitative microbial risk assessment. Wiley, New York
56. McBride G, Till D, Ryan T, Balll A, Lewis G, Palmer S, Weinstein P (2002) Pathogen occurrence and human health risk assessment analysis. In: Freshwater Microbiology Research Programme, Ministry for the Environment, Ministry of Health, New Zealand
57. Soller JA (2009) Potential implications of person-to-person transmission of viral infection to US EPA's Groundwater Rule. J Water Health 7(2):208–223
58. Soller JA, Eisenberg JNS (2008) An evaluation of parsimony for microbial risk assessment models. Environmetrics 19(1): 61–78
59. Eisenberg JNS, Seto EYW, Olivieri AW, Spear RC (1996) Quantifying water pathogen risk in an epidemiological framework. Risk Anal 16(4):549–563
60. Soller JA, Eisenberg J, DeGeorge J, Cooper R, Tchobanoglous G, Olivieri A (2006) A public health evaluation of recreational water impairment. J Water Health 4(1):1–19
61. Soller JA, Olivieri AW, Crook J, Cooper RC, Tchobanoglous G, Parkin RT, Spear RC, Eisenberg JN (2003) Risk-based approach to evaluate the public health benefit of additional wastewater treatment. Environ Sci Technol 37(9):1882
62. Soller JA, Bartrand T, Ashbolt NJ, Ravenscroft J, Wade TJ (2010) Estimating the primary aetiologic agents in recreational freshwaters impacted by human sources of faecal contamination. Water Res 44(16):4736–4747
63. Wong M, Kumar L, Jenkins TM, Xagoraraki I, Phanikumar MS, Rose JB (2009) Evaluation of public health risks at recreational beaches in Lake Michigan via detection of enteric viruses and a human-specific bacteriological marker. Water Res 43(4): 1137–1149
64. Ashbolt A, Reidy C, Haas CN (1997) In Microbial health risk at Sydney's coastal bathing beaches. In: 17th Australian Water and Wastewater Association meeting, AWWA, Melbourne, 1997, pp 104–111
65. Ashbolt N, Bruno M (2003) Application and refinement of the WHO risk framework for recreational waters in Sydney, Australia. J Water Health 1(3):125–131
66. Roser D, Ashbolt N (2007) Source water quality assessment and the management of pathogens in surface catchments and aquifers. Research report 29, Bolivar
67. Schoen ME, Ashbolt NJ (2010) Assessing pathogen risk to swimmers at non-sewage impacted recreational beaches. Environ Sci Technol 44(7):2286–2291
68. Steyn M, Jagals P, Genthe B (2004) Assessment of microbial infection risks posed by ingestion of water during domestic water use and full-contact recreation in a mid-southern African region. Water Sci Technol 50(1):301–308
69. Diallo MB, Anceno AJ, Tawatsupa B, Houpt ER, Wangsuphachart V, Shipin OV (2008) Infection risk assessment of diarrhea-related pathogens in a tropical canal network. Sci Total Environ 407(1):223–232
70. Gerba CP, Rose JB, Haas CN, Crabtree KD (1996) Waterborne rotavirus: a risk assessment. Water Res 30(12):2929–2940
71. Kay D, Fleisher JM, Salmon RL, Jones F, Wyer WD, Godfree AF, Zelenauch-Jacquotte Z, Shore R (1994) Predicting likelihood of gastroenteritis from sea bathing: results from randomised exposure. Lancet 344:905–909
72. Wade TJ, Calderon RL, Brenner KP, Sams E, Beach M, Haugland R, Wymer L, Dufour AP (2008) High sensitivity of children to swimming-associated gastrointestinal illness: results using a rapid assay of recreational water quality. Epidemiology 19(3):375–383
73. Wade TJ, Calderon RL, Sams E, Beach M, Brenner KP, Williams AH, Dufour AP (2006) Rapidly measured indicators of recreational water quality are predictive of swimming-associated gastrointestinal illness. Environ Health Perspect 114(1):24–28
74. Soller JA, Schoen ME, Bartrand T, Ravenscroft J, Ashbolt NJ (2010) Estimated Human Health Risks from Exposure to Recreational Waters Impacted by Human and Non-human Sources of Faecal Contamination. Water Res 44(10):4674–4691
75. Teunis PF, Moe CL, Liu PSEM, Lindesmith L, Baric RS, Le Pendu J, Calderon RL (2008) Norwalk virus: how infectious is it? J Med Virol 80(8):1468–1476
76. Eisenberg JNS, Soller JA, Scott J, Eisenberg DM, Colford JMJ (2004) A dynamic model to assess microbial health risks associated with beneficial uses of biosolids. Risk Anal 24(1):221–236
77. Eisenberg JNS, Moore K, Soller JA, Eisenberg D, Colford JMJ (2008) Microbial risk assessment framework for exposure to amended sludge projects. Environ Health Perspect 116(6):727–733
78. Eisenberg JNS, Seto EYW, Colford JM, Olivieri AW, Spear RC (1998) An analysis of the Milwaukee cryptosporidiosis outbreak based on a dynamic model of the infection process. Epidemiol 9(3):255–263
79. Vikesland PJ, Wigginton KR (2010) Nanomaterial enabled biosensors for pathogen monitoring - A review. Environ Sci Technol 44(10):3656–3669
80. Goodwin KD, Litaker RW (2008) Emerging Technologies fir monitoring recreational waters for bacteria and viruses. In: Walsh PJ, Smith SL, Fleming LE, Solo-Gabriele HM, Gerwick WH (eds) Oceans and human health. Academic, San Diego
81. Boehm AB, Yamahara KM, Love DC, Peterson BM, McNeill K, Nelson KL (2009) Covariation and photoinactivation of traditional and novel indicator organisms and human viruses at a sewage – impacted marine beach. Environ Sci Technol 43:8046–8052
82. Gersberg RM, Rose MA, Robles-Sikisaka R, Dhar AK (2006) Quantitative detection of hepatitis A virus and enteroviruses near the United States-Mexico border and correlation with levels of fecal indicator bacteria. Appl Environ Microbiol 72(12):7438–7444
83. Griffin DW, Gibson CJ, Lipp EK, Riley K, Paul JH, Rose JB (1999) Detection of viral pathogens by reverse transcriptase

PCR and of microbial indicators by standard methods in the canals of the Florida Keys. Appl Environ Microbiol 65(9): 4118–4125

84. Aulicino FA, Orsini P, Carere M, Mastrantonio A (2001) Bacteriological and virological quality of seawater bathing areas along the Tyrrhenian coast. Int J Environ Health Res 11:5–11
85. Alonso JL, Alonso MA (1993) Presence of Campylobacter in marine waters of Valencia. Spain Water Res 27(10):1559–1562
86. Gonzalez AM, Paranhos R, Lutterbach MS (2010) Relationships between fecal indicators and pathogenic microorganisms in a tropical lagoon in Rio de Janeiro, Brazil. Environ Monit Assess 164(1–4):207–219
87. Lang AL, Tsai Y-L, Mayer CL, Patton KC, Palmer CJ (1994) Multiplex PCR for Detection of the Heat-Labile Toxin Gene and Shiga-Like Toxin I and II Genes in Escherichia coli Isolated from Natural Waters. Appl Environ Microbiol 60(9): 3145–3149
88. Graczyk TK, Sunderland D, Tamang L, Lucy FE, Breysse PN (2007) Bather density and levels of Cryptosporidium, Giardia, and pathogenic microsporidian spores in recreational bathing water. Parasitol Res 101(6):1729–1731
89. Alexander LM, Heaven A, Tennant A, Morris R (1992) Symptomatology of children in contact with sea water contaminated with sewage. J Epidemiol Community Health 46(4):340–344
90. Bandaranayake DR, Turner SJ, McBride GB, Lewis G, Till D (1995) Health effects of bathing at selected New Zealand marine beaches. New Zealand Department of Health, Wellington
91. Fewtrell L, Godfree AF, Jones F, Kay D, Salmon RL, Wyer MD (1992) Health effects of white-water canoeing. Lancet 339:1587–1589
92. Fewtrell L, Kay D, Salmon R, Wyer M, Newman G, Bowering G (1994) The health effects of low-contact water activities in fresh and estuarine waters. J Inst Water Environ Manag 8:97–101
93. Fleisher JM, Kay D, Salmon RL, Jones F, Wyer MD, Godfree AF (1996) Marine waters contaminated with domestic sewage: nonenteric illnesses associated with bather exposure in the United Kingdom. Am J Public Health 86(9):1228–1234
94. Haile RW, Witte JS, Gold M, Cressey R, McGee C, Millikan RC, Glasser A, Harawa N, Ervin C, Harmon P, Harper J, Dermand J, Alamillo J, Barrett K, Nides M, Wang G.-Y (1996) An epidemiological study of possible adverse health effects of swimming in Santa Monica Bay. Santa Monica Bay Restoration Project, Santa Monica, CA, 70pp
95. Kueh CSW, Tam TY, Lee T, Wong SL, Lloyd OL, Yu ITS, Wong TW, Tam JS, Bassett DCJ (1995) Epidemiological study of swimming-associated illnesses relating to bathing-beach water quality. Water Sci Technol 31:1–4
96. Lee JV, Dawson SR, Ward S, Surman SB, Neal KR (1997) Bacteriophages are a better indicator of illness rates than bacteria amongst users of a white water course fed by a lowland river. Water Sci Technol 35:165–170
97. Lightfoot N (1989) A prospective study of swimming related illness at six freshwater beaches in southern Ontario. University of Toronto, Toronto
98. Marino F, Morinigo M, Martinez-Manzanares E, Borrego J (1995) Microbiological-epidemiological study of selected marine beaches in Malaga (Spain). Water Sci Technol 31:5–9
99. Medema G, Van Asperen I, Klokman-Houweling J, Nooitgedagt A, Van de Laar M (1995) The relationship between health effects in triathletes and microbiological quality of freshwater. Water Sci Technol 31:19–26
100. Medical Research Council & Council for Scientific and Industrial Research (CSIR) (1995) Pathogenic microorganisms/ epidemiological microbiological study. Final report (1991–1995), South Africa
101. Philipp R, Evans EJ, Hughes AO, Grisdale SK, Enticott RG, Jephcott AE (1985) Health risks of snorkel swimming in untreated water. Int J Epidemiol 14(4):624–627
102. Pike E (1994) Health effects of sea bathing (WMI 9021), phase III—final report to the department of the environment. Water Research Centre, Medmenham
103. Prieto M, Lopez B, Juanes J, Revilla J, Llorca J, Delgado-Rodriguez M (2001) Recreation in coastal waters: health risks associated with bathing in sea water. J Epidemiol Community Health 55(6):442–447
104. Stevenson AH (1953) Studies of bathing water quality and health. Am J Public Health Nations Health 43:529–538
105. United Nations Environment Programme/World Health Organization (1991) Epidemiological studies related to environmental quality criteria for bathing waters. Shellfish-growing waters and edible marine organisms (Activity D). Final report on project on relationship between microbial quality of coastal seawater and rotavirus-lnduced gastroenteritis among bathers (1986–88). United Nations Environment Programme, Athens, Greece
106. United Nations Environment Programme/World Health Organization (1991) Epidemiological studies related to environmental quality criteria for bathing waters, Shellfish-growing waters and edible marine organisms (Activity D). Final report on epidemiological study on bathers from selected beaches in Malaga, Spain (1988–1989). United Nations Environment Programme, Athens, Greece
107. van Asperen IA, Medema G, Borgdorff MW, Sprenger MJ, Havelaar AH (1998) Risk of gastroenteritis among triathletes in relation to faecal pollution of fresh waters. Int J Epidemiol 27(2):309–315
108. van Dijk PAH, Lacey RF, Pike EB (1996) Health effects of sea bathing—further analysis of data from UK beach surveys. Department of the Environment, WRc pic, Medmenham
109. von Schirnding YE, Kfir R, Cabelli V, Franklin L, Joubert G (1992) Morbidity among bathers exposed to polluted seawater. A prospective epidemiological study. S Afr Med J 81(11):543–546

Recycling Collection and Materials Separation

Matthew J. Franchetti
Department of Mechanical, Industrial, and Manufacturing Engineering, University of Toledo, Toledo, OH, USA

Article Outline

Glossary
Definition of the Subject and Its Importance
Introduction
The Hierarchy of Solid Waste Management
Benefits from Recycling
Fundamentals of Recycling Processes
Common Problems and the Human Factors of Recycling
Solid Waste and Recycling Assessments
Recycling Collection Systems and Methods
Separation Methods and Equipment by Material
Financial Analysis
Case Study
Industrial Ecology and Solid Waste Exchanges
Future Directions
Bibliography

Glossary

Source reduction Source reduction is the reduction of materials coming into the system. The US Environmental Protection Agency (EPA) defines source reduction as "activities designed to reduce the volume or toxicity of waste generated, including the design and manufacture of products with minimum toxic content, minimum volume of material, and/or a longer useful life." The two interesting components of this definition are *volume* and *toxicity*. This means that an organization does not have to solely focus on reducing volumes for source reduction initiatives, but could focus efforts on reducing the negative impacts on the environment of those same volumes. This may mean a reduction of pallets entering a building or of flash being generated from a mold.

Reuse Reuse is the actual reuse of a material in its present form. Some examples are printing draft copies on the back side of previously used paper, using incoming pallets as outgoing pallets, or using incoming boxes as collection containers for recyclables.

Recycling Recycling is a type of reuse which involves changing the composition or properties of the material in one way or another. For example, this can be accomplished by melting it down, chipping, or grinding the materials. A broad definition of recycling is taking a product or material at the end of its useful life and turning it into a usable raw material to make another product. There are three types of recycling, as discussed in the hierarchy of waste management: in-process, on-site, and off-site.

Pollution prevention Pollution prevention is the broadest and most difficult term to concisely define. In essence, it is the overall process of reducing waste and preventing pollution from entering the environment through the air, water, or ground. It encompasses both the aspects of source reduction and waste reduction. The EPA has defined pollution prevention as follows: Pollution prevention means source reduction, as defined under the Pollution Prevention Act, and other practices that reduce or eliminate the creation of pollutants through: Increased efficiency in the use of raw materials, energy, water, or other resources, or Protection of natural resources under conservation.

Definition of the Subject and Its Importance

A broad definition of recycling is taking a product or material at the end of its useful life and transforming it into a usable raw material to make another product. Recycling collection and separation is concerned with the infrastructure, resources, and processes required for gathering, transporting, separating, and consolidating materials for the recycling process. Recycling collection and material separation are critical components of solid waste minimization and aid in conserving natural resources, maintaining healthier environments, and reducing greenhouse gas emissions. In addition, many organizations can achieve economic benefits and public image enhancements by developing strong recycling programs.

Introduction

In the year 2006, US residents, businesses, and institutions generated more than 228 million metric tons (251 million US tons) of solid waste, which is approximately 2.1 kg (4.6 pounds) of waste per person per day, up from 1.2 kg (2.7 pounds) per person per day in 1960 [1]. From this total, 32.5% is recovered and recycled or composted, 12.5% is burned at combustion facilities, and the remaining 55% is disposed of in landfills [1]. Source reduction can be a successful method of reducing waste generation. Practices such as grass recycling, backyard composting, two-sided copying of paper, and transport packaging reduction by industry have yielded substantial benefits through source reduction. Source reduction has many environmental benefits. It prevents emissions of many greenhouse gases, reduces pollutants, saves energy, conserves resources, and reduces the need for new landfills and combustors.

Recycling, including composting, diverted 74 million metric tons (82 million US tons) of material away from disposal in 2006, up from less than 14 million metric tons (15 million US tons) in 1980, when the recycle rate was just 10% and 90% of MSW was being combusted with energy recovery or disposed of at landfills [1]. Typical materials that are recycled include automobile batteries, recycled at a rate of 99%, paper and paperboard at 52%, and yard trimmings at 62%. These materials and others may be recycled through curbside programs, drop-off centers, buyback programs, and deposit systems.

International Waste Generation Comparison Rates

As discussed in the book "Germany, Garbage, and the Green Dot" by Bette Fishbein, it is very difficult to make international comparisons regarding waste generation [2]. For example, Ms. Fishbein points out that "according to the data published by the Organization of Economic Cooperation and Development, waste generation in Germany is 318 kg per person per year as compared to 864 kg per person per year in the United States. This might suggest that the average person in the United States generates two or three times as much garbage as the average person in Germany. However, data from the two countries are not comparable: the German data do not include materials collected from recycling, nor do they include some commercial waste, both of which are included in U.S. data. International comparisons of waste generation are usually unreliable because countries use different data collection methodologies and different definitions of waste."

Recycling is a critical component to help increase these material recovery rates. This section provides an overview of the hierarchy of solid waste management and the benefits of recycling. Next, an overview of the fundamentals of recycling processes is discussed, along with common problems and human factors of recycling systems. Accurate data are critical when developing recycling systems, and a brief overview of solid waste and recycling assessments is provided. Then, a discussion of collection methods and separations equipment is provided, followed by a discussion regarding the financial analyses related to recycling process decisions. A case study is also provided to demonstrate these analyses. Finally, an overview of industrial ecology and waste exchanges is provided and a discussion of future directions in this field.

The Hierarchy of Solid Waste Management

When approaching solid waste analysis and minimization, there is an accepted hierarchy, or order of solution approaches, that should be deployed when analyzing an organization's waste streams. This hierarchy is given before a discussion of definitions of solid waste management terms because it makes the definitions more meaningful and easier to understand. This hierarchy has been defined in the US Pollution Prevention Act of 1990 as follows:

The Congress hereby declares it to be national policy of the United States that pollution should be prevented or reduced at the source whenever feasible; pollution that cannot be prevented should be recycled in an environmentally safe manner, whenever feasible; pollution that cannot be prevented or recycled should be treated in an environmentally safe manner whenever feasible; and disposal or release into the environment should be employed only as a last resort and should be conducted in an environmentally safe manner.

Based on that definition, the solid waste management solutions, in order of preference, are:

1. Source reduction
2. In-process recycling
3. On-site recycling

4. Off-site recycling
5. Waste treatment to render the waste less hazardous
6. Secure disposal
7. Direct release to the environment

These terms may be easier to understand with an example. A very straightforward and simple example involves a company that receives incoming raw material in cardboard packaging. Suppose the company is a metal stamping plant for the automobile industry. Below are situations where the company could apply the hierarchy:

1. Source reduction – eliminate the waste through engineering process modifications. For example, the company could work with the vendor to eliminate the use of cardboard containers and switch to returnable plastic containers. This is the most preferred method and completely eliminates the problem of waste management and need for recycling.
2. In-process recycling – if waste is generated, develop a separation method to use the waste as a raw material for the same process. For example, the company could collect the metal scrap from the stamping process and attempt to process it again in the same process.
3. On-site recycling – if waste is generated, develop a separation method to use the waste as a raw material for another in-house process. For example, the company could use the cardboard packages to ship their own final products.
4. Off-site recycling – if waste is generated, develop a separation method and transport the waste to another organization so that another company could use the waste as a raw material. For example, the company could transport the cardboard to a third-party processor for recycling.
5. Waste treatment to render the waste less hazardous – if waste is generated, develop a separation method and treat the waste so it is less harmful before releasing it to the environment. This solution applies mostly to chemical processes, such as treating waste water and releasing it to the public sewer.
6. Secure disposal – dispose of the waste at a secure landfill. The company could have their waste hauler transport the cardboard to a local landfill.
7. Direct release to the environment – when waste is generated, develop a separation method and release the waste directly to the environment. In this final scenario, the company could stage the cardboard in an outside area to allow it to biodegrade.

Benefits from Recycling

The purpose of a project, plan, or initiative (solid waste minimization or otherwise) is to achieve measurable results that can be tied into the original goal. These results or benefits are often critical in determining the feasibility or acceptance of a project proposal. These benefits are also the key selling points used when promoting solid waste minimization to stakeholders and decision makers. The benefits of solid waste minimization and recycling can be separated into four areas:

- Environmental
- Economic
- Corporate image
- Personal and social

Ideally, an organization would like to create a situation where multiple benefits can be realized from a single project. This synergistic approach allows for the creation of win-win situations when applied appropriately using the system approach discussed in this book. Specifically, the company will realize cost benefits and enhanced public image, the environment will be protected, and the stakeholders of the organization (including employees) often gain a sense of well-being and harmony with the environment as the organization is protecting the greater good for society. This chapter discusses in greater detail these benefits and includes examples that may be used to promote solid waste minimization to decision makers.

Environmental Benefits

Waste minimization efforts are a big step forward in moving toward a sustainable environment. The results are clear: cleaner air and water, less pollution, more forested land and open space, and reduced greenhouse gases. It is obvious that recycling translates into less trash entering landfills. But the greatest environmental benefits of recycling are not related landfills, but to the conservation of energy and natural resources and the prevention of pollution when a recycled material, rather than a raw material, is used to make a new product. Since recycled materials have been refined and processed once, manufacturing the second time around is much cleaner and less energy intensive than the first. The following list summarizes the key benefits to the environment that can be derived from solid waste minimization:

- Conservation of natural resources (water, trees, energy, and land)
- Healthier environment via landfill emissions reductions (carbon dioxide, methane, and leachate)
- Global warming reduction
- Conservation of habitats

The primary environmental benefit of solid waste minimization is resource conservation. The Medical University of South Carolina (MUSC) reports that the college recycled 1,151 metric tons (1,269 US tons) of paper, metals, organics, and other materials in 2003. Based on the school's calculations, this saved a total of about 14,513 kJ (13,756 BTU) of energy, enough energy to power nearly 137 homes for 1 year. In addition, products made using recovered rather than virgin or raw materials use significantly less energy. Less energy used means less burning of fossil fuels such as coal, oil, and natural gas. When burned, these fuels release pollutants, such as sulfur dioxide, nitrogen oxide, and carbon monoxide, into the air. By using recycled materials instead of trees, metal ores, minerals, oil, and other raw materials harvested from the earth, recycling-based manufacturing conserves the world's scarce natural resources. This conservation reduces pressure to expand forests cutting and mining operations.

Recycling and composting in the USA diverted nearly 63.5 million metric tons (70 million US tons) of material away from landfills and incinerators in 2000 as reported by the National Recycling Coalition. This total is up from 31 million metric tons (34 million US tons) in 1990, which is doubling in just 10 years. Below are some interesting facts about the relationship between recycling and resource conservation. These facts can have great emotional appeal when promoting waste minimization and can serve as part of a comprehensive strategy to promote recycling.

- Every US ton of paper that is recycled saves 17 trees.
- The energy that is saved when one glass bottle is recycled is enough to light a light bulb for 4 h.
- Recycling benefits the air and water by creating a net reduction in ten major categories of air pollutants and eight major categories of water pollutants.
- In the USA, processing minerals contributes almost half of all reported toxic emissions from industry, sending 1.4 million metric tons of pollution into the air and water each year. Recycling can significantly reduce these emissions.
- It is important to reduce our reliance on foreign oil. Recycling helps the nation accomplish this by saving energy.
- Manufacturing with recycled materials, with very few exceptions, saves energy and water and produces less air and water pollution than manufacturing with virgin materials.
- It takes 95% less energy to recycle aluminum than it does to make it from raw materials. Making recycled steel saves 60%, recycled newspaper 40%, recycled plastics 70%, and recycled glass 40%. These savings far outweigh the energy created as by-products of incineration and landfilling.
- In 2000, recycling resulted in an annual energy savings equal to the amount of energy used in six million homes (over 696 trillion kJ [660 trillion BTU]). In 2005, recycling is conservatively projected to save the amount of energy used in nine million homes (950 trillion kJ [900 trillion BTU]).
- A national recycling rate of 30% reduces greenhouse gas emissions as much as removing nearly 25 million cars from the road.
- Recycling conserves natural resources, such as timber, water, and minerals.
- Every bit of recycling makes a difference. For example, 1 year of recycling on just one college campus, Stanford University, saved the equivalent of 33,913 trees and the need for 577 metric tons (636 US tons) of iron ore, coal, and limestone.
- When one US ton of steel is recycled, 1,134 kg (2,500 pounds) of iron ore, 635 kg (1,400 pounds) of coal, and 54.4 kg (120 pounds) of limestone are conserved.
- Brutal wars over natural resources, including timber and minerals, have killed or displaced more than 20 million people and are raising at least $12 billion a year for rebels, warlords, and repressive governments. Recycling eases the demand for the resources.

- Mining is the world's most deadly occupation. On average, 40 mine workers are killed on the job each day, and many more are injured. Recycling reduces the need for mining.
- Tree farms and reclaimed mines are not ecologically equivalent to natural forests and ecosystems.
- Recycling prevents habitat destruction, loss of biodiversity, and soil erosion associated with logging and mining.

Solid waste minimization also aids in creating a healthier environment by reducing landfill emissions. As discussed in the environmental concerns section of this chapter, landfills emit a liquid called leachate. Leachate is liquid that is generated from a landfill that is created from decomposing waste, created after rainwater mixes with the chemical waste in a landfill, or liquids present in the landfill. Once it enters the environment, the leachate is at risk for mixing with groundwater near the site which can have very negative effects. This liquid can be treated in a similar manner to sewage, and the treated water can then be safely released into the environment. One study reports that landfills are responsible for 3.8% of the global warming damage from human sources in the USA. Municipal solid waste landfills are the largest source of human-related methane emissions in the United States, accounting for about 25% of these emissions in 2004 [1]. This gas consists of about 50% methane (CH_4), the primary component of natural gas; about 50% carbon dioxide (CO_2); and a small amount of nonmethane organic compounds. In 2003, US landfills generated 131.2 teragrams methane in terms of carbon dioxide (CO_2) equivalents (where a teragram is 1 million metric tons). Reducing the amounts of solid waste disposed in landfills would reduce methane generation and subsequently reduce global warming. Although it is difficult to accurately quantify habitat loss, many animal species are displaced by the creation of landfills and the effects of deforestation. By minimizing solid waste levels and increasing recycling, available habitats for animals will not be disrupted by the development or expansion of landfills and the effects of deforestation to acquire virgin raw materials.

The Medical University of South Carolina reported that in 2003, recycling reduced overall air emissions by 22.6 metric tons (24.9 US tons) (excluding carbon dioxide and methane) and reduced waterborne waste by 3.8 metric tons (4.2 US tons). By reducing air and water pollution and saving energy, recycling offers an important environmental benefit: It reduces emissions of greenhouse gases, such as carbon dioxide, methane, nitrous oxide, and chlorofluorocarbons, that contribute to global climate change. Recycling and composting reduce greenhouse gas by:

- Decreasing the energy needed to make products from raw materials
- Reducing emissions from incinerators and landfills, which are the largest source of methane gas emissions in the USA
- Slowing the harvest of trees, thereby maintaining the carbon dioxide storage benefit provided by forests

Recycling prevents the emission of many greenhouse gases and water pollutants, saves energy, supplies valuable raw materials to industry, creates jobs, stimulates the development of greener technologies, conserves resources for our children's future, and reduces the need for new landfills and combustors.

Recycling also helps reduce greenhouse gas emissions that affect global climate. In 1996, recycling of solid waste in the United States prevented the release of 30 million metric tons (33 million US tons) of carbon into the air – roughly the amount emitted annually by 25 million cars. The number of landfills in the United States is steadily decreasing – from 8,000 in 1988 to 1,754 in 2006. The capacity, however, has remained relatively constant. New landfills are much larger than in the past.

Economic Benefits

The economic benefits of waste minimization are often one of the key selling points when promoting environmentally conscious initiatives to businesses. Other than regulatory compliance, the cost benefits from waste minimization can turn an "environmental decision" into a wise business decision that will increase an organization's financial statements. Often times, when promoting a solid waste minimization program to the decision makers of an organization, the most influential benefits are the cost savings and potential revenue generated from the program. Often, when the creation of a recycling program is first discussed with management, the first response is "we do not have a budget for recycling." This is far from the truth; the budget does exist, and the starting point is the funds that the company is currently paying for waste hauling and removal. The systems approach to solid waste minimization explores for cost-effective methods to better utilize these funds and protect the environment. The three areas for cost benefits are usually derived from:

- Cost avoidance in solid waste hauling and disposal
- Cost savings in material purchases due to reuse and reduction
- Revenue generation from the sale of recyclable material

Many organizations are surprised to learn that recycling and waste minimization can make strong business sense. A common environmental adage is "become green to make green." The Business Waste Reduction Assistance Program at The University of Toledo has identified over $3.1 million in annual savings for Northwest Ohio businesses in the 70 waste assessments that the program has completed. For example, at a plastic manufacturer with 100 employees, approximately $16,000 in annual cost benefits were identified via increased plastic, paper, and cardboard recycling. The waste stream amounts and revenues also cost justified the purchase of a baling system.

In the state of Pennsylvania, recycling adds significant value to the state's economy. In the state, collection and processing, the first step in the recycling process, involves sorting and aggregating recyclable materials. It includes municipal and private collectors, material recovery and composting facilities, and recyclable material wholesalers. These activities employ nearly 10,000 people in Pennsylvania, with a payroll of $284 million and annual sales of $2.3 billion. Recycling manufacturing involves the actual conversion of recyclables into products. The primary recycling manufacturers in Pennsylvania in order of magnitude are steel mills, plastic converters, paper and paperboard mills, and nonferrous metal manufacturers. Recycling manufacturing employs over 64,000 people with a payroll of almost $2.5 billion and annual sales of over $15.5 billion. Reuse and remanufacturing focuses on the refurbishing and repair of products to be reused in their original form. The largest activities are retail sales of used merchandise and reuse of used motor vehicle parts. The amount of value that can be added via this process is limited because of competition from new products. Nevertheless, reuse and manufacturing contributes over 7,000 jobs, a payroll of $115 million, and sales of over half a billion dollars.

On a national scale, the recycling industry continues to grow at a rate greater than that of the economy as a whole. According to the Institute for Local Self-Reliance, total employment in the recycling industry from 1967 to 2000 grew by 8.3% annually while total United States employment during the same period grew by only 2.1% annually. The recycling industry also outperformed several major industrial sectors in regard to gross annual sales as its sales rose by 12.7% annually during this period. Furthermore, the number of recycling industries in the United States increased from 8,000 in 1967 to 56,000 in 2000. These facilities employ 1.1 million people across the country.

For many items, recycling can be more cost effective versus disposal. The list below provides a summary of select construction materials based on survey results of 63 companies:

- Average cost to recycle
  - Asphalt debris: $5.70 per US ton
  - Concrete rubble: $4.85 per US ton
  - Used bricks and blocks: $5.49 per US ton
  - Trees and stumps: $37.69 per US ton
  - Wood scrap: $46.43 per US ton
- Average cost of disposal
  - Over $75.00 per US ton and can be as high as $98.00 per US ton

Recycling saves money for manufacturers by reducing energy costs. In 2001, New Jersey's recycling efforts saved a total of 135 kJ (128 trillion BTU) of energy, equal to nearly 17.2% of all energy used by industry in the state, with a value of $570 million.

The sale of recycled products is an increasingly important component of the retail sector and commerce in general. There are over 1,000 different types of recycled products on the market, and due to changes in technology and increased demand, today's recycled products meet the highest quality standards. Recycled products are also more readily available than ever before and are affordable. By purchasing recycled products, consumers are helping to create long-term stable markets for the recyclable materials that are collected from New Jersey homes, businesses, and institutions.

The economic value of clean air, water, and land is significant, but difficult to quantify. Since recycling plays an important role in protecting these natural resources, it must be attributed an economic value in this context as well.

Corporate Image Benefits

Corporate imaging and product branding play a critical role in the profitability of any organization. Successfully maintaining and strengthening these concepts is one of the chief duties of any marketing department, and environmental initiatives can go a long way to bolster them. Specifically, by focusing on solid waste minimization and publicizing these efforts, an organization can:

- Increase sales by attracting environmentally conscious consumers
- Improve the recruitment of employees that share similar values
- Attract environmentally conscious partners
- Attain free corporate publicity
- Increase employee involvement gateway to other programs (heart and mind)
- Maintain cleaner facilities

Personal and Social Benefits

Solid waste minimization also offers personal and social benefits. Although many of these benefits are somewhat intangible and difficult to measure, they are worth mentioning. They are worth mentioning because they can be selling points when promoting an environmental program. Below is a list of some of these benefits:

- Personal satisfaction for helping the environment
- Sustainable environment for future generations
- Cleaner facilities
- Buy-in at work programs (employee involvement)
- Healthier environments and a higher standard of living
- Generation of money to assist local programs such as the sale of aluminum cans to benefit a children's burn unit at a hospital

Fundamentals of Recycling Processes

An understanding of the recycling industry and the related processes can be very beneficial when evaluating recycling system options. In addition, a basic understanding can aid managers and engineers in making better decisions in regard to recycling and disposal options. Following is a brief summary of some of these benefits:

1. Gaining a better understanding of recycling options and the different roles that material recovery facilities, processors, and material brokers play in the process
2. Gaining an understanding of the recycling process and the process flow for materials as they leave a facility
3. Gaining an understanding of the material separation needs based on the recycling process for each material type
4. Gaining an understanding of the recycling process to design better processes and products to reduce the environmental impact
5. Learning more about the LCA process and develop more accurate inventory audits

This section provides an overview of the recycling industry which includes brief discussions of the business entities operating in the field. In addition, overviews of the recycling processes for major waste items are also provided. The intent of this section is to expose the reader to basic terms and processes in the field, not to provide a detailed or comprehensive analysis.

Recycling Industry Overview

The recycling industry is comprised of five primary entities that work together to get recyclable materials from the point of generation as a by-product or waste to the stage where they can be used again as raw materials.

These entities are:

- Haulers – these companies transport materials between entities, including the generation facility, consolidation points, and processing facilities. Often times, these companies will lease a semitruck trailer to the generating facilities to consolidate and store recyclable materials before transportation to a consolidator or depot.

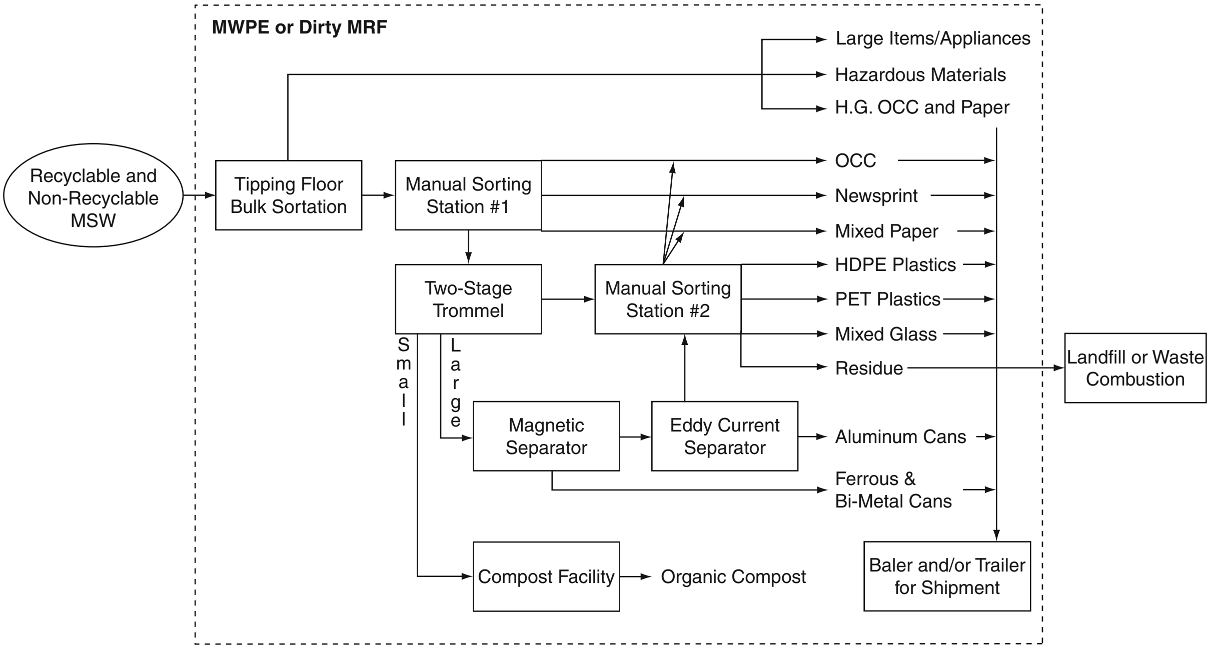
- Material recovery facilities (MRF) – an MRF is a specialized plant that receives, separates, and prepares recyclable materials for marketing to end-user manufacturers [3]. There are two types of MRFs: clean and dirty MRFs. A clean MRF accepts recyclable materials that have been collected as commingled wastes from curbside collection or drop-off sites (Fig. 2). The most common clean MRF is a two-stream MRF, where source-separated recyclables are delivered in the form of a mixed commingled container stream (typically glass, ferrous metal, aluminum and other nonferrous metals, PET [No. 1] and HDPE [No. 2] plastics) and a mixed paper stream. A dirty MRF accepts a mixed (recyclable and nonrecyclable) solid waste stream and then proceeds to separate recyclable materials via a combination of manual and mechanical sorting [3] (Fig. 1). The sorted recyclable materials may undergo further processing required to meet technical specifications established by end markets while the balance of the mixed waste stream is sent to a disposal facility such as a landfill [3].
- Consolidators and depots – a consolidator or depot is similar to an MRF, but it does not perform any sorting operations. These entities hold or store materials until a specified batch size is reached or when a recycling processing facility is ready to process the material.
- Material brokers – material brokers buy recyclable materials from cities, businesses, depots, or MRFs and sell the materials to a processing facility.
- Processing facilities – these facilities perform the actual processes to recycle materials. Many different processing facilities exist for different materials, such as metals, glass, and papers.

Aluminum Recycling

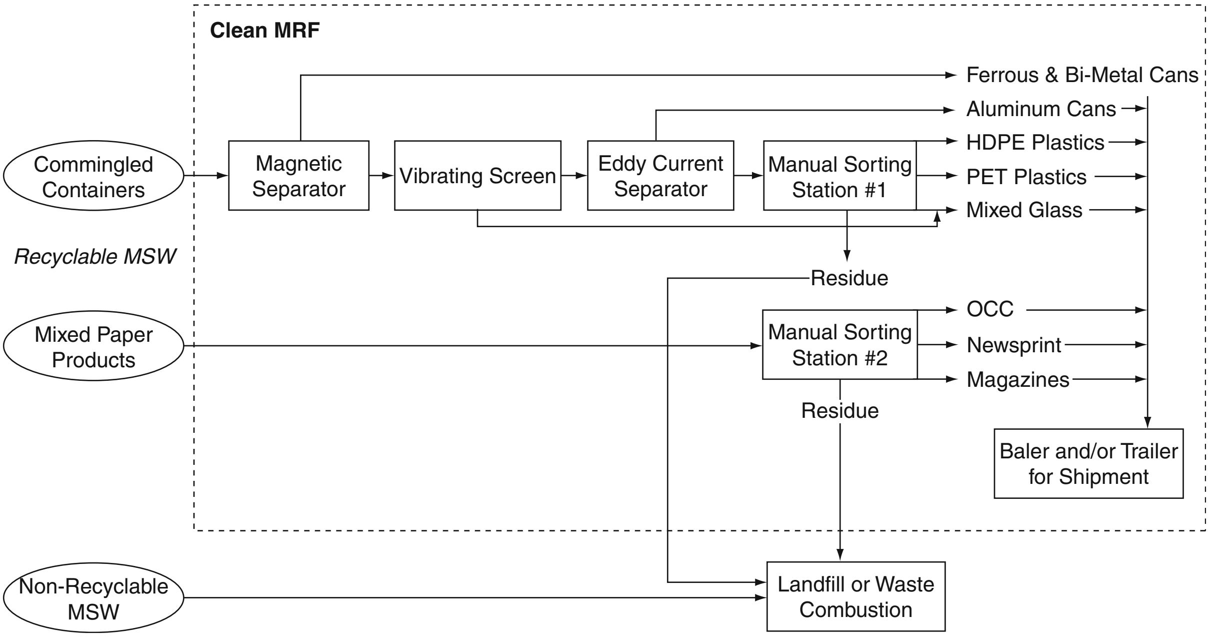
Aluminum recycling can be broken down into three steps:

- Sorting
- Baling
- Compressing

A United Kingdom Web site (rethink.sita.co.uk) provides an overview of aluminum recycling and is summarized in the following paragraphs [4]:



Recycling Collection and Materials Separation. Figure 1
Dirty MRF overview



Recycling Collection and Materials Separation. Figure 2 Clean MRF overview

Before being transported to an aluminum recycling facility, the material is passed under a magnet to remove any steel. Aluminum is a non-ferrous metal, so it is not magnetic. Steel is a ferrous metal, so it is magnetic. The magnet picks up the steel cans and separates them from aluminum. The remaining cans are then crushed and compressed to form bales and transported to the aluminum recycling facility.

Once at the facility, the cans are shredded into small pieces about 2.5 cm in diameter. These pieces then pass through a magnet which removes any remaining ferrous metals such as steel. The shredded aluminum then travels to a de-coater, where hot air (at 500°C/930°F) removes any coating or decoration.

The hot cans go straight from the de-coater to the furnace, where they are melted at a temperature of 700°C/1,300°F. Once melted, the liquid aluminum is transferred to a holding furnace which clears any remaining contaminants, and a degasser which removes any gas.

The liquid aluminum is poured into cooled rectangular shaped moulds. The cooling transforms the aluminum back into a solid metal. This solid metal is taken to a saw where the ends are made square and transported to the rolling facility. At the rolling facility the aluminum is rolled into large sheets to be used a raw material.

Glass Recycling

Different types of glass go through different recycling processes. For example, cookware melts at a much higher temperature than container glass and must be processed separately. This section follows the typical recycling process of container glass (such as beverage bottles). There are four types of glass as related to recycling processes [5]:

- Container glass (wine and beer bottles)
- Float glass (windows)
- Cookware (plates and dishes)
- Automotive glass (windshields)

Glass for recycling is mostly collected from businesses, curbside collection, or community drop-off sites. Trucks collect the bottles and transport them to be stored at a depot. When a processing batch of glass has been collected and delivered to the depot, it is all transported to a glass recycling facility [5].

Once at the recycling facility, the glass is crushed (crushed glass is "cullet"). Cullet goes through various processes to remove nonglass items. To remove ferrous metal, the cullet is passed through a strong magnet which removes the ferrous metals such as steel and iron [5]. To remove nonferrous metals, the cullet passes by powerful air jets which separate the metal pieces from the cullet [5]. To remove lightweight items, such as paper, the cullet goes through a vacuum [5]. To remove any remaining items that are not glass, such as ceramics, the cullet passes under a laser which rejects them [5]. The cullet is now ready to be made into new glass. To make new glass, the cullet goes into a furnace where it is melted at a temperature of 1,500°C/2,700°F [5]. The high temperature turns the cullet into a liquid called molten glass. The molten glass is shaped into molds to become bottles or jars. Recycled glass is melted at a lower temperature than virgin glass, which saves 30% of the energy used [5].

Paper Recycling

Collected paper must be sorted before being recycled. The recycling process itself can be separated into eight steps [6]:

- Sorting
- Baling
- Pulping
- Screening
- De-inking
- Pouring
- Rolling
- Packing

This section provides a brief overview of each of these steps. Paper recycling can be challenging because there are over 50 grades of wastepaper. The main four groups are [6]:

- Low grade (mixed paper, corrugated board)
- De-inking grade (newspapers, magazines, office paper)
- Kraft grade (unbleached brown backing)
- High grade (printer cutoffs and unprinted paper)

Large amounts of paper, including shredded paper, are baled before being transported to a paper mill. Once at the paper mill, the paper is placed into a large vat and mixed with water. The process breaks down the paper into tiny strands of cellulose fibers. Eventually, this turns into a mushy mixture called pulp.

The pulp is then filtered and screened. The screens are made of a series of holes and slots of different shapes and sizes, and remove any remaining contaminants such as bits of plastic of glue. For certain uses, pulp must also be de-inked. There are two main methods of de-inking [6]:

1. Washing – Chemicals can be used to separate the ink from the paper, and the ink is washed away with water. Although this process requires the use of chemicals and water, the quantities used are much less than the manufacture of new paper and the water can often be cleaned and reused.
2. Floatation – Air can be passed through the pulp to produce foam. The foam holds at least half of the ink and can be skimmed off.

Pulp is poured into a huge flat wire screen. On the screen, water starts to drain from the pulp and the recycled fibers quickly begin to bond together to form a watery sheet. The sheet, which now resembles paper, passes through a series of heavy rollers, which squeeze out more water; some heated cylinders, which dry the paper; and an iron roller, which irons the paper [6]. Next, the paper is wound into a large roll. One roll can be as wide as 9 m (30 ft) and weigh as much as 18 metric tons (20 US tons). The roll of paper is cut into smaller rolls, or sometimes sheets, before being dispatched for use.

Plastic Recycling

Plastic recycling can be separated into six steps [7]:

- Sorting
- Shredding
- Cleaning
- Melting
- Extrusion
- Pelletizing

Plastics are synthetic polymers made from oil and natural gas and are one of the world's most used raw materials [7]. Plastics are blended in different formulas and modified with additives to create the 40 categories of plastic and the several specific grades within these formulas [7]. Most consumer plastics in waste are labeled with an Identification Code, numbered 1–7. Before plastics are processed, they are sorted into seven different polymer types. The polymer type indicates both the properties and characteristics of the material, such as the melt temperature and its suitability for recycling [7]. The symbols used to classify the different polymer types can be found on the plastic items as follows:

Number 1: PETE or PET (polyethylene terephthalate)

- Used typically for beverage bottles, plastic wrap, and frozen food trays.
- Recycled PETE is commonly used to make new clothing fibers, furniture, new beverage containers, and carpet.

Number 2: HDPE (high-density polyethylene)

- Used typically for milk jugs, juice jugs, liquid detergent bottles, trash and shopping bags, and cereal box liners.
- Recycled HDPE is commonly used to make detergent, oil and vitamin bottles, drain pipes, recycling bins, dog houses, and other plastic lumber.

Number 3: Vinyl (PVC)

- Used in clear food packaging, shampoo bottles, and medical tubing. It is also used in construction building.
- Recycled PVC is used to make packaging, binders, mats, decks, paneling, roadway gutters, mud flaps, and speed bumps.

Number 4: LDPE (low-density polyethylene)

- Used in food packaging (especially bread), frozen food bags, shrink wrap, dry cleaning bags, and squeezable bottles.
- Recycled LDPE is used to make mailing envelopes, trash cans, trash liners, furniture, floor tiles, paneling, and compost bins and lumber.

Number 5: PP (polypropylene)

- Used to make ketchup and medicine bottles, some dairy containers, and molded automobile parts.
- Recycled PP is commonly used to make signal lights, car battery components, brooms, brushes, oil funnels, ice scrapers, bike racks, pallets, storage bins, and trays.

Number 6: PS (polystyrene)

- Used to make packing foam, egg cartons, meat trays, aspirin bottles, plates, CD jackets, and food service items.
- Recycled PS is used to make thermometers, switch plates, insulation, egg cartons, vents, office supplies, foam packaging, and containers.

Number 7: Other plastics

- Typical examples are 3- and 5-gallon reusable water bottles, some citrus juice and ketchup bottles.

Once sorted, the plastics are baled before being transported to a plastic reprocessing plant. Once at the reprocessing plant, the plastic is shredded into small pieces which are then washed. After washing, the plastic pieces are passed under a metal detector to remove any metal, and a de-dusting unit which removes any lighter particles [7].

The clean plastic pieces are dried and melted so they can be made into new shapes. The melted plastic is then filtered to remove any remaining contaminants and extruded to form fine strands. The plastic strands are then cut into pellets, cooled in water, then dried and stored ready to be processed and molded as new plastic items [7].

Common Problems and the Human Factors of Recycling

There are many common problems involved in creating a successful recycling program. The two biggest problems found in the duration of this research were material misplacement and hindrances inherent in the company itself.

Material Misplacement

One problem often encountered when companies attempt to separate out recyclables is employee involvement. At most sites surveyed, containers dedicated to certain materials were found to contain other materials as well. This often results in higher costs for the companies. For instance, at one company, recyclable material was found in a dumpster dedicated for hazardous wastes and therefore costs more to dispose. However, wood, Old Corrugated Containers (OCC), and other recyclables were finding their way into this dumpster.

At another company, recyclables were found in the landfilled trash stream even though there were recycling containers nearby. One such example was in a warehouse. The warehouse had recycling containers for shrink wrap and clean cullet (broken glass) positioned throughout the aisles, and yet the shrink wrap, glass, and wood were all found in the garbage cans which were also throughout the aisles. In addition, there were aluminum collection barrels containing glass, glass collection barrels containing aluminum, and other miscellaneous trash mixed into the bins labeled for recyclables.

A third company, a heavy manufacturer and assembly plant, also encountered this problem of material misplacement. Upon a preliminary tour of the plant, it was noted that in clearly marked recycling containers, other materials such as expanded polystyrene cups, paper, and mixed trash were found. Similarly, in a large office building with over 600 employees, the paper collection boxes also contained film plastic, OCC, and mixed trash.

These observations raised several serious questions. First, why do people place the wrong materials into collection containers or place recyclables into a landfill stream even when the appropriate container is located within close proximity? What is the motivation which causes some people to obey the recycling policies and what is the motivation that causes others not to? What can be done to improve involvement and deter carelessness in regard to recycling? What human factors come into play in determining the signs which should be used to mark collection containers and in the placement of containers? Finally, what motivational ideas work best in regard to recycling, trash separation, and waste minimization?

It has been noted that when recycling containers are distinctive according to acceptable material, the employees are much more likely to place the proper material in the container. This may involve a difference in the type of container used, the color of the recycling container, or the size and shape of the opening or of the container itself. For instance, some companies use large wire bins which are always for recyclables only. Other used colored containers to designate the end point of the materials. Still others use containers of different shapes and sizes for each different material. Often, the same types of similar containers are used with only a difference in size and shape of the opening for each material.

It is also obvious that when there are clear signs designating the acceptable materials for each container, accidental wrong placement of materials is much less likely. These signs usually work better if they are in color, have pictures, and are placed in easy sight. The signs must be easily recognizable and distinguishable in order to facilitate employee involvement. People do not like to be bothered by extra time and energy needed in order to recycle. If it requires extra effort, people will often not participate.

In a similar venue, the collection containers must be properly placed or people will not bother to find them. Individuals who have a personal interest in recycling will hold on to recyclables until they can be properly disposed of or placed in the proper container. Such people have often been found to do such things as take aluminum cans home from work if the company does not recycle them. These people will take the added time and energy necessary to get the materials in their proper containers. However, if one does not have such personal convictions, then in fast-paced society, taking the extra time and effort to find the proper recycling container is often too time-consuming and problematic to bother with.

If there is a motivation to recycle, the outcome is much better. There is not one cure-all for motivating employees, but there are some highly successful motivation factors. These include money, fun, and free time. There three motivation factors can be utilized in a variety of ways. One option is to keep track of recycling by department and then to award the best department(s) with cash, a party, or a paid afternoon off. There are countless other creative options available. The feasibility of these options depends on the particular company and its schedule, policies, and recycling revenue.

According to human nature, there will always be some people who are too careless, apathetic, or lazy to obey recycling mandates. However, the proper use of human factors and motivation can minimize the number of such people and in turn minimize the incidents of contamination of recyclables and recycling collection containers.

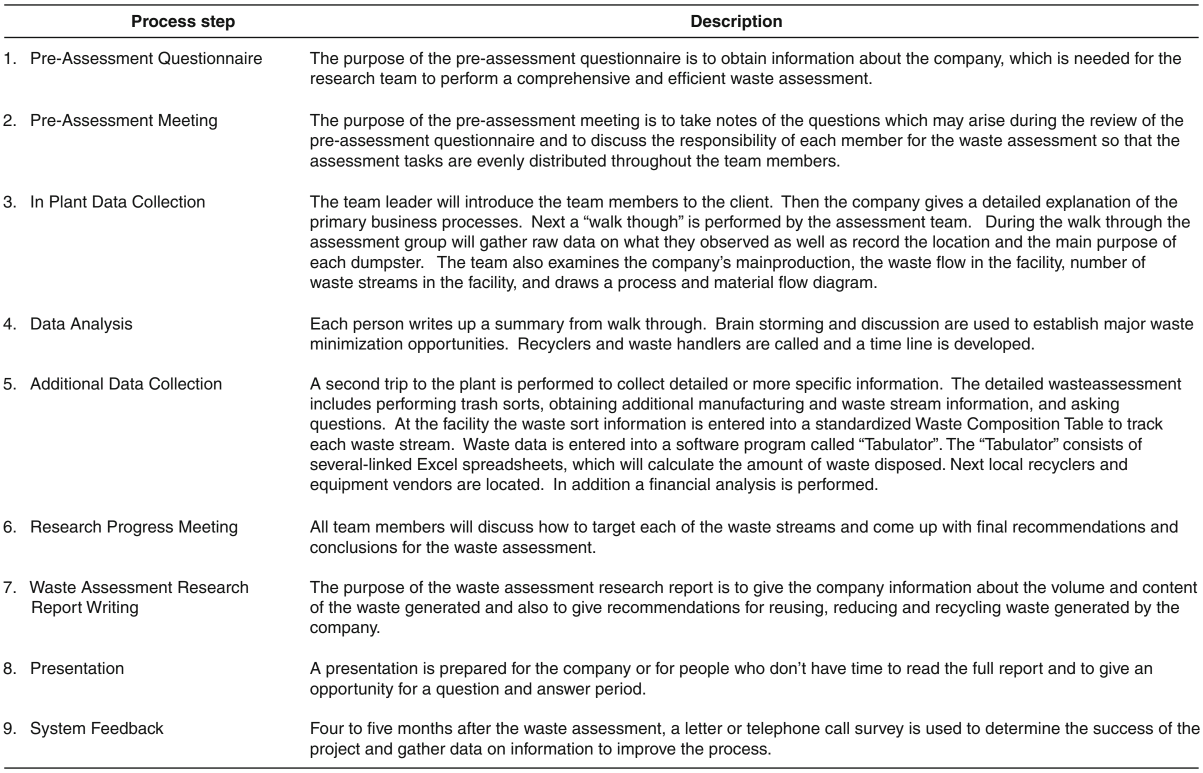
Observed Common Hindrances

There are many common hindrances to recycling which occur on a higher level than employee involvement. These include management perceptions, company policies, union rules and regulations, poor past performance in recycling attempts, and many other reasons. These hindrances must be overcome or successful recycling will be impossible. It is often very difficult to overcome these hindrances. Very often, they are due to a misunderstanding or wrong perceptions. This makes it vitally important for the assessment team to understand the hindrances and how to combat them.

Though it may seem impossible to overcome policies or company rules, recommendations may still be made which do not conform to the problematic rules. The recommendations can be provided as win-win situations and may go a long way to adjusting the policies inhibiting recycling practices. The most important issue is that the economically and ecologically best scenario is chosen. This can always be accomplished with win-win situations if those participating are willing to be creative. In short, the assessment team cannot direct the company to change its policies but can present alternatives which allow the company, management, or union to see the downfall in the policy and the benefits of alterations.

Solid Waste and Recycling Assessments

The process used for the assessment was expanded from a waste assessment manual provided by the US EPA [8]. This EPA manual was also used by Petek to reduce waste water emissions by 24% at a textile facility in Slovenia [9]. To initiate the waste audit process, the company was given a Pre-assessment Questionnaire to complete. This is done to allow the researchers the opportunity to have a general understanding of the company before gathering data on the walk-through. The questionnaire gives the researchers an opportunity to investigate the processes beforehand if needed. Figure 3 summarizes the nine-step waste assessment process.



Process step	Description
1. Pre-Assessment Questionnaire	The purpose of the pre-assessment questionnaire is to obtain information about the company, which is needed for the research team to perform a comprehensive and efficient waste assessment.
2. Pre-Assessment Meeting	The purpose of the pre-assessment meeting is to take notes of the questions which may arise during the review of the pre-assessment questionnaire and to discuss the responsibility of each member for the waste assessment so that the assessment tasks are evenly distributed throughout the team members.
3. In Plant Data Collection	The team leader will introduce the team members to the client. Then the company gives a detailed explanation of the primary business processes. Next a "walk though" is performed by the assessment team. During the walk through the assessment group will gather raw data on what they observed as well as record the location and the main purpose of each dumpster. The team also examines the company's mainproduction, the waste flow in the facility, number of waste streams in the facility, and draws a process and material flow diagram.
4. Data Analysis	Each person writes up a summary from walk through. Brain storming and discussion are used to establish major waste minimization opportunities. Recyclers and waste handlers are called and a time line is developed.
5. Additional Data Collection	A second trip to the plant is performed to collect detailed or more specific information. The detailed wasteassessment includes performing trash sorts, obtaining additional manufacturing and waste stream information, and asking questions. At the facility the waste sort information is entered into a standardized Waste Composition Table to track each waste stream. Waste data is entered into a software program called "Tabulator". The "Tabulator" consists of several-linked Excel spreadsheets, which will calculate the amount of waste disposed. Next local recyclers and equipment vendors are located. In addition a financial analysis is performed.
6. Research Progress Meeting	All team members will discuss how to target each of the waste streams and come up with final recommendations and conclusions for the waste assessment.
7. Waste Assessment Research Report Writing	The purpose of the waste assessment research report is to give the company information about the volume and content of the waste generated and also to give recommendations for reusing, reducing and recycling waste generated by the company.
8. Presentation	A presentation is prepared for the company or for people who don't have time to read the full report and to give an opportunity for a question and answer period.
9. System Feedback	Four to five months after the waste assessment, a letter or telephone call survey is used to determine the success of the project and gather data on information to improve the process.

Recycling Collection and Materials Separation. Figure 3
Waste assessment process overview

Recycling Collection Systems and Methods

Advances in technology, changes in recycling laws, trends in recycling markets, and other variables have had a large impact on recycling collection systems and methods. The public and businesses now have more options to aid them in increasing recycling levels. So, it is important for organizations to periodically evaluate their recyclables collection system to determine if it is the most cost-effective and efficient program. Several widely used programs include:

- Curbside programs (single stream, multiple stream, and pay-as-you-throw)
- Automated collection
- Drop-off centers
- Incentive programs

Curbside Recycling

Curbside recycling now serves half of the US population, providing the most convenient means for households to recycle a variety of materials [10]. While individual curbside programs differ, the most commonly collected materials are "The Big Five," which includes aluminum cans, glass bottles, paper, plastic, and steel/tin cans. Curbside recycling exists in several ways [10]. They include:

- Dual-stream recycling
- Single-stream recycling
- Pay-as-you-throw

Dual-stream recycling is probably the most popular form of curbside recycling in the USA [11]. Used containers (plastics and metals) are placed into one bin, and papers (such as newspaper, magazines, and direct mail) go in another bin. Both bins are set out on the curb on pickup day. Most communities that offer this service use special trucks divided into two compartments so workers can sort at the truck.

Single-Stream Recycling is a method that is growing, but somewhat controversial. It provides one cart (65 or 94 gal) where materials are commingled [12]. Households do not have to separate any materials. Haulers favor single-stream because it involves less trucks and pickups [12]. Evidence suggests that single-stream increases the quantity of household recyclables, and many cities have implemented single-stream programs as a result [12].

Pay-as-you-throw (PAYT) is a trash collection program that encourages curbside recycling [13]. Residents are charged per trash bag, and curbside recycling is offered at no or a reduced cost. There are several benefits to PAYT programs as discussed by the EPA [13]:

- Decreases waste: The US EPA says municipalities often see 25–35% less nonrecyclable waste.
- Increases recycling: If residents can pay for trash or recycle for "free," they are much more watchful about what gets trashed; one California PAYT program saw recycling volumes triple, literally overnight.
- Control of waste costs: Residents have a direct effect on what they spend on disposal.
- More information about who supports PAYT is available from the US EPA. Over 5,000 communities across the country currently have a PAYT program.

Automated Collection

Many communities are moving toward automated collection using a specialized vehicle that lifts, empties, and returns a cart to the curb. The driver controls the entire process from the cab of the vehicle and does not leave the vehicle [10]. This can reduce labor costs and on-the-job injuries, thereby reducing worker's compensation claims [10]. Automated systems utilize specialized trucks equipped with mechanical extensions that automatically lift and empty trash and recycling containers without the driver leaving the vehicle. This is a system designed to improve the efficiency and makes the task of putting out garbage easier and cleaner for the residents [10].

Drop-off Centers

Drop-off centers can be a cost-effective way for smaller or rural communities to collect recyclables [10]. It can also be effective for urban communities to offer businesses, apartment buildings, and condominiums access to recycling [10]. In most cases, drop-off centers have lower operating costs versus curbside collection due to lower resource requirements in terms of labor and fuel.

Drop-off centers, like curbside programs, can offer full source separation (multiple bins), commingled materials (fewer bins), or a single-stream (fully commingled) approach for collecting recyclables [10].

Dual-Stream Versus Single-Stream Collection

Single-stream recycling collection: "Single-stream" recycling collection (fully commingled) programs allow participants to put all recyclable materials (e.g., paper, bottles, cans) into one collection container. In the case of paper, all grades are mixed together along with cardboard and rinsed food containers at the curb and/or bin at a municipal transfer station.

The fully commingled recyclable materials are then transferred to a central point such as a materials recovery facility (MRF) where the recyclables are separated using handpicking and/or automation. Paper collected in single-stream systems may be further separated into various paper grades (e.g., high-grade white paper, newsprint, cardboard). For single-stream recycling to work, the processing facility must sort the recyclable materials properly and thoroughly to meet market specifications [14].

Single-stream collection has become popular for the following reasons:

- Many residents find it convenient, since no sortation is required by them.
- The convenience has resulted in higher recycling participation rates [3].
- New materials can be added to the collection system with minimal change to the collection system and process.
- It is found to increase the amount of recyclables collected [3].
- Reduces the number of collection trucks needed because single-stream programs usually lead to changing to automated or semiautomated curbside collection [3].

Dual-stream or multi-stream recycling collection programs require participants to place each recyclable material in the appropriate collection bin when they first discard the item. Separate containers collect glass, metal, plastic, newsprint, and magazines.

Recovered paper can be collected separately by grade (e.g., white office paper, newspapers, magazines, and corrugated cardboard boxes) or, more commonly, collected as mixed paper separated from other recyclable materials. If different grades of paper are commingled, they are sorted at a central point, such as a materials recovery facility (MRF). If paper is separated at the source, the different grades of paper can be marketed separately for the highest return.

Benefits of dual- or multi-stream (sorted) collection include [3]:

- Lower levels of contamination at the source.
- Higher-quality materials.
- Materials are more valuable and may result in higher financial returns.
- Lower costs to process the recovered paper.

Single-stream recycling collection allows residents to "fully commingle" all their recyclables (mixed paper and mixed food containers together) at the curb, transfer station, or recycling center [10]. The materials are separated during processing, at a resource recovery or materials recovery facility, and prepared for recycling markets [10]. Single-stream curbside collection programs are most often implemented with automated or semiautomated collection [10].

Incentive Programs

The following is a list of several incentive programs that have been implemented to increase recycling levels for communities:

"Recycle Man" Rewards Residents with Grocery Gift Cards [10]

Polk County, Florida – Officials recently conducted a 6-week recycling incentive program. Each recycling day, the county recycling coordinator, also known as "the Recycle Man" and his supervisor canvassed local communities with low recycling rates to encourage residents to recycle. For each bin they observed that was properly prepared, they awarded the resident with a $20 gift card to their local grocery store. While full results are pending, the county did receive a notable increase in calls for bins as a result, and their MRF reported an increase in recycling volume [10].

Recyclebank [10]

RecycleBank provides coupons to residents that participate in curbside recycling programs.

RecycleBank calculates and records quantity (weight) of recyclables collected at curb (requires automated collection using bar code on recycling container) in exchange for coupons for products and services for local businesses. In 2008, the City of Hartford, CT implemented a 1-year pilot program that included using RecycleBank [10].

Door Prizes Encourage Participation at County Drop-off Center [10]

Hanover, Pennsylvania – Hanover, Pennsylvania, was ordered by the PA Department of Environmental Protection to make curbside recycling mandatory for all residents in the summer of 2004. The city immediately began looking for ways to track and promote recycling participation [10].

One method officials used was to issue a card with a bar code to each resident subscribed to the recycling service. Residents swipe the card when they take recyclables to the recycling center. The cards help the borough keep track of residents' participation in the recycling program [10].

To create incentive for residents to continue dropping off recycled materials at the center, the borough began issuing door prizes at random to people using the recycling center starting in May 2006. More than 800 cards were swiped at the recycling center during the first 2 weeks of March [10].

Separation Methods and Equipment by Material

A materials recovery facility (MRF) or materials reclamation is a specialized plant that receives, separates, and prepares recyclable materials for marketing to end-user manufacturers. As previously discussed, there are two different types – clean and dirty MRFs.

The percentage of residuals (unrecoverable recyclable or nonprogram materials) from a properly operated clean MRF supported by an effective public outreach and education program should not exceed 10% by weight of the total delivered stream, and in many cases, it can be significantly below 5% [3]. A dirty MRF recovers between 5% and 45% of the incoming material as recyclables [3]; then, the remainder is landfilled or otherwise disposed. A dirty MRF can be capable of higher recovery rates than a clean MRF, since it ensures that 100% of the waste stream is subjected to the sorting process, and can target a greater number of materials for recovery than can usually be accommodated by sorting at the source [3]. However, the dirty MRF process is necessarily labor intensive, and a facility that accepts mixed solid waste is usually more challenging and more expensive to site [3]. The following sections provide additional details regarding the recycling processes for common materials.

Newspaper

The generalized process for newspaper collection involves:

- Newspaper is collected and consolidated from drop-off sites, curbside collection, or business collection.
- The material is baled at a material recovery facility.
- The baled material is sold to a paper mill.
- The paper mill recycles the newspaper and creates raw material to be used as newspapers or other products.

Old newspaper is an essential material in the paper remanufacturing process; as the paper mills are concerned about both quality (cleanliness, type of paper) and quantity of the supply, they usually issue purchasing contracts to dealers rather than buying small amounts of paper from the public [15].

The equipment and recycling process are summarized in the following paragraph:

At the paper mill, de-inking facilities separate ink from the newspaper fibers through a chemical washing process. A slusher turns the old paper into pulp, and detergent dissolves and carries the ink away. Next, screens remove contaminants like bits of tape or dirt. The remaining pulp is bleached and mixed with additional pulp from wood chips to strengthen it. The watery mixture is poured onto a wire, a continuously moving belt screen which allows excess moisture to drain through. By the time the mixtures gets to the end of the belt, it's solid enough to be lifted off and fed through steam-heated rollers which further dry and flatten it into a continuous sheet of paper. This paper machine produces finished newsprint at the rate of 915 m (3,000 ft) per minute [15].

Cardboard

The generalized process for cardboard or Old Corrugated Container (OCC) recycling is very similar to newspaper recycling. The equipment and recycling process are summarized in the following paragraph:

While some corrugated cardboard is recycled at curbside, the bulk of it comes from commercial rather than residential sources. At the paper mill, the corrugated containers are pulped and blended with additional pulp from wood chips. Old fibers, which are broken, shorter and weaker, as compared to new fibers, are blended with the new pulp. Recycled paper fibers and new pulp are then blended to make linerboard. Next, the linerboards are shipped to a boxboard plant, where the manufacturing process is finished. The paper is corrugated by specially-geared machines, the linerboards are glued on, and the resulting flat pieces, called mats, are trimmed to size and creased along a pattern of folds. The mats are shipped flat to customers who set them up into boxes. Then the boxes are used to package products for shipping [15].

Glass

The generalized process for glass recycling also involves collection and consolidation. Glass recycling poses additional challenges over other materials because it is fragile and breaks easily during transportation and consolidation. This can cause safety and separation issues for transportation and hauling companies. The equipment and recycling process are summarized in the following paragraph:

At the plant, a mechanical processing system breaks the glass into small pieces called cullet. Magnets, screens and vacuum systems separate out metals, labels, bits of plastic, metal rings and caps. The cullet then is blended in measured amounts with silica sand, soda ash, and limestone, and placed in a furnace which melts it into molten glass [10].

Steel (Tin) Cans

The equipment and recycling process are summarized in the following paragraphs:

After the cans are collected, the volume of cans collected and type of transportation arrangements available will determine whether the load will go through a dealer or directly to a de-tinning plant. At the plant, a chemical de-tinning solution flows into and drains the cans more easily, which results in better recovery of the tin during the reclaiming process. The process consists of a series of chemical and electrical steps which separate, purify, and recover the steel and tin. In the batch process of de-tinning, the cans first are loaded into perforated steel drums and dipped into a caustic chemical solution which dissolves the tin from the steel. The de-tinned steel cans are drained, rinsed, and baled into 14”x14”x30” 181 kg (400 pound) blocks. Then they are sold to steel mills to be made into new products [10].

Meanwhile, the liquid with the tin, a salt solution called sodium stannate, is filtered to remove scraps of paper and garbage. Then it’s chemically treated to eliminate other metals. Next, the solution is transferred to an electrolysis bath. When electricity is applied, tin forms on one of the plates in the solution. After the plate is covered, the tin is melted off and cast into ingots. The ingots are at least 99.98% pure tin and are used in the chemical and pharmaceutical industries. Pure tin also is alloyed with other metals to make solder, babbitt, pewter, and bronze products [10].

Aluminum

The equipment and recycling process are summarized in the following paragraph:

After the aluminum is collected and consolidated it is transported to a smelter, where it may be shredded or ground into small chips before being melted and cast into ingots. The ingots are sent on to manufacturing plants where they are rolled into sheets of aluminum and used to manufacture end products ranging from cans to castings to car bodies. The major markets for shredded aluminum are exports (comprising a variety of end users) and domestic smelters [10].

Financial Analysis

If this section were to be reduced into one key idea, it would be the application of the three R’s and the two E’s as they apply to solid waste minimization. From a technical and regulatory standpoint, the hierarch mentioned in the previous section is excellent, but it does not hit home with the leadership of many organizations and does not promote the full benefits of solid waste minimization [16]. In terms of communicating and promoting solid waste minimization, the three R’s and the two E’s have served as a very effective tool. The three R’s are:

- *Reduce*
- *Reuse*
- *Recycle*

The three R's are a summary of the hierarchy discussed in the previous section. In terms of the hierarchy, the first item discussed, "source reduction," has been separated into two components, "reduce" and "reuse." Reuse has been added to emphasize the fact that many items that are being disposed of at a landfill by organizations could have been reused, such as cardboard containers, plastic caps, or rubber bands [16]. Finally, all recycling methods, in-process, on-site, and off-site, have been lumped into one category for simplicity.

And the two E's are:

- *E*nvironment
- *E*conomics

The concept is to apply the three R's at an organization to help the two E's. The three R's provide the solutions to the solid waste problem based on the hierarchy of solid waste management. The two E's communicate the goals of these efforts: to lessen the environmental impact of an organization and improve an organization's economics or bottom line. This simple phrase is easy to understand and has served as a great "tag line" or "catch phrase" to promote solid waste reduction. By emphasizing economic benefits, solid waste reduction and minimization benefits are much easier to sell to the decision makers because the efforts are placing attention on the financial health of the organization as well. A plant manager of a battery manufacturing firm located in Toledo, Ohio, bluntly summarized this concept when he stated that his company "is in the business of producing and selling batteries, not recycling" [16]. It is easy to get caught up in the good feelings associated with helping the environment, but unless there is a financial incentive, many organizations may give environmental concerns only "lip service" [16]. A buzzword that is now emerging is "green washing," which is a situation in which a company publicly and verbally promotes their environmental efforts to bolster corporate images, but falls short of the actions associated with the public statements. The goal of emphasizing the three R's and two E's is to bridge the gap between public statements and corporate actions by demonstrating that environmental concerns make business sense.

A common misconception of organizations is that they do not have funds in their budgets for recycling or environmental initiatives. The fallacy in this thinking is that virtually all companies have a starting point or a "budget" for recycling and waste reduction: the annual expenses for trash removal and janitorial efforts [16]. The paradigm shift requires considering the amount of money expended per year as the "environmental" or "recycling" budget and devising creative methods and processes to minimize impacts to the environment. Many organizational leaders are surprised to learn the potential cost avoidance or revenue generated from becoming more environmentally conscious. A financial case study is discussed later in this section highlighting some of the typical benefits.

Economic benefits from solid waste analysis and minimization can be achieved a variety of ways. The most common economic benefits derived from solid waste minimization are listed below [16]:

- Cost avoidance – Organizations can save money by diverting solid waste streams from the landfill to back within the company in terms of reuse or off-site to a recycler. The monetary savings are derived from no longer paying a waste hauler to remove the trash and dispose of it at the landfill.
- Recycling revenue – Substantial additional revenue can be earned by selling recyclables to third-party processors or recycling commodity brokers. For example, one US ton of baled cardboard sells for $100–$180 on the market.
- Reduced raw material costs – When an organization is able to utilize in-process or on-site recycling, they reduce their raw material needs directly by replacing virgin material purchases with in-house scrap, rework, or process by-products.
- Reduced energy costs – By reducing the amount of materials within a facility through reduction and reuse, material handling costs can be minimized.
- Increased sales – Many consumers and businesses look favorably on organizations that are environmentally conscious and purchase products or services from them.
- Increased productivity – As workers are engaged in efforts that they see as meaningful, many of them take pride and put additional efforts into their work. In addition, absenteeism may reduce as well.

The case study discussed in the following section highlights the financial aspects of a decision by the

Lucas County Solid Waste Management District to build and operate a government-owned material recovery facility.

Case Study

Solid waste minimization and recycling goals for municipalities are achievable through the installation of material recovery facilities (MRFs), and in certain solid waste management systems, government-owned and government-operated MRFs are feasible and cost justified. This section demonstrates a process to evaluate and determine the operational and economic feasibly of a government-owned MRF [17]. A case study from Lucas County, Ohio, is provided that demonstrated this analysis process. The key findings from the case study indicate that the municipality will achieve a payback period of approximately 4 years, and a 10-year internal rate of return of 20.5%, versus the current system of outsourcing [17].

In 2007, the Lucas County Solid Waste Management District (District) purchased a material recovery facility (MRF) to sort and sell nearly 9,000 metric tons (10,000 US tons) recyclable materials that were collected per year from its municipal recycling programs. Currently, the county recycles two classes of materials, commingled fiber and commingled used beverage containers. In 2006, the District recycled 9,755 US tons of materials [17]: 7,280 US tons were commingled fiber (mixed office paper, newspaper, and OCC) and 2,475 were commingled used beverage containers (plastic bottles and aluminum cans).

The first phase of the analysis process involved estimating the current recycling levels in terms of materials compositions and volumes (annual tonnages). These data were collected from District records from the 2006 fiscal year and included operating cost and revenue data [17]. Once combined, this information provided a complete baseline of the operations of the current system utilizing outsourced processing. This baseline was used to compare the cost structure of acquiring a county-owned and county-operated MRF. The baseline data all annualized costs and revenues associated with the drop-off recycling program, specifically [17]:

- Revenue paid from third-party processors for recyclable materials
- Third-party processing fees
- Labor costs
- Administrative costs
- Vehicle costs (fuel, maintenance, repair)
- Drop-off container and material costs

The second phase involved indentifying potential MRF sites. A local business realtor was contacted for assistance. Upon the identification of the MRF site, a complete annual cost and revenue projection was conducted to operate the MRF over a 20-year period [17]. This analysis included the following annualized costs and revenues:

- Revenue paid from third-party recycling material commodity brokers
- Building purchase cost (including realtor fees)
- Building modification and renovation costs
- Equipment and inspection/repair costs
- Labor costs (including driver and processors)
- Administrative costs
- Utility costs
- Vehicle costs (fuel, maintenance, repair)
- Drop-off container and material costs

This financial projection of the proposed MRF was compared with the current system baseline. In essence, the analysis answered the question whether the additional revenue earned from the sale of the processed recyclable materials will outweigh the additional capital and operating costs over the projected 20-year life of the project at a 15% minimum attractive rate of return [17]. To accomplish this analysis, a net present worth (NPW) was conducted. The NPW is the difference between the present worth of all cash inflows and outflows of a project. Since all cash flows are discounted to the present, the NPW method is also known as the discounted cash flow technique. This method allows not only the selection of a single project based on the NPW value but also a selection of the most economical project from a list of more than one alternative projects, in the case of this research, the existing system of outsourcing versus purchasing and operating a government-owned MRF.

This section demonstrated the process for municipalities to economically justify the purchase and operation of a government-owned MRF. Key findings from this research revolve around a case study from the 2007 purchase of a government-owned MRF in Toledo, Ohio, USA [17]. The key findings were demonstrated through a complete financial analysis. Specifically, the financial analysis indicated that the municipality will achieve a payback period of approximately 4 years and a 10-year internal rate of return of 20.5% [17]. The consequences of these findings, stemming from the economic and operational justification, led to the actual purchase of the MRF site and subsequent operation in 2007 through early 2008. This research may serve as an example or model for other local governments considering the implementation of such a system [17].

Industrial Ecology and Solid Waste Exchanges

Industrial ecology is the field of research that studies waste generation from a macrolevel for entire industries. From a solid waste standpoint, industrial ecology is concerned with the conversion or reuse of undesirable materials into something useful for another company or industry, in other words, waste exchanges and material efficiency. Material efficiency is defined as the percentage of process by-products that are recycled or reused divided by the total by-product generation for a company or industry. Industrial ecology and waste exchanges examine the material efficiency and methods to improve that efficiency. Whereas waste audits examine an individual company's ability to reduce, reuse, or recycle, waste exchanges examine an entire industry's or region's ability to reduce, reuse, or recycle. In essence, waste exchanges examine methods for one company to use another company's by-products as a raw material, diverting this material from entering a landfill. Waste exchanges are a great tool that can enhance a company's recycling levels and generate economic benefits as part of the solid waste auditing process. Waste streams identified during a solid waste audit that the company cannot reduce or reuse could be sent to another company using one of many solid waste exchanges operating around the world.

With increased pressures on companies to improve profitability and reduce environmental impacts, waste exchanges are more popular than ever. Many companies and nonprofit organizations are turning to these exchanges to bolster corporate images and reduce costs. The Internet has simplified, streamlined, and reduced the costs associated with the administration of waste exchanges as well. Information is available in real time, 24 h/day, which makes such systems more accurate and user friendly, while allowing the exchanges to reach a larger client base.

History and Background

Waste commodity exchange is defined as the ability of a company or organization to use another company's waste as its raw material. As the old adage goes, "One person‘s trash is another person's treasure." Instead of sending seemingly worthless items or process by-products away to a landfill, the goal of the waste commodity exchange is to find a company that may get more use out of the product.

A good household example of this is garage sales, which are an excellent way to reuse products. Another alternative is to find different ways to reuse items. Baby food jars, for example, can be reused to store miscellaneous nuts, bolts, and washers in a workshop.

Waste exchanges have been around for over 60 years [18]. The British Government established the earliest documented industrial waste exchange, called the National Industrial Materials Recovery Association, in 1942 [18]. This waste exchange was created to conserve materials for the war effort during World War II. The first North American waste exchange was started in Canada in 1974 for hazardous waste [18]. The National Industrial Materials Recovery Association is no longer active as it disbanded after the war. The Canadian waste exchange is still active as the Canadian Waste Materials Exchange (CWME).

Waste commodity exchanges are reuse and recycling services that help these types of material exchanges occur on a much larger scale for businesses. These services help businesses save money as well as help the environment by diverting waste into usable raw materials.

Waste Commodity Exchanges in North America

Over 200 waste commodity exchanges are currently operating in North America. These exchanges differ in terms of the service area, materials exchanged, exchange processes, and fee structures [19]. Many of these exchanges are coordinated by state and local governments, while others are for-profit businesses. The US Environmental Protection Agency (Washington) provides an excellent reference list of waste exchanges and contact information [19].

More than 35 national and 150 state-specific waste exchanges exist in the USA, and Canada has more than seven national waste exchanges [19]. The majority of the waste exchanges are specific to certain regions or states. The drawback to regional or state-specific exchanges is that they expose the available materials to fewer potential companies. The benefits of regional exchanges, though, are that they significantly reduce transportation costs, especially for heavy or bulky items and large quantities.

Regional exchanges are appealing to companies that may continually exchange waste items over an extended period of time due to longstanding process by-products. An example of this is plastic scrap from a manufacturing process. Another company may be able to grind the scrap, use it as a raw material, and establish dedicated routes to transport the material. On the other hand, national exchanges expose materials to a much larger number of companies, but transportation fees may make some options infeasible.

The material and waste focus of the various exchanges differs significantly. Some are very broad and deal with a wide variety of materials. For example, in terms of the national exchanges, Recycler's World (www.recycle.net) and the Reuse Development Organization (www.redo.org) handle any waste that users post on the Web site.

Alternatively, some exchanges are very narrowly focused. Good national examples of this are the American Plastics Exchange (www.apexq.com), which deals solely with plastics, and Planet Salvage (www.planetsalvage.com), which deals only with used automobile parts. Overall, any material that is available from one business and wanted by another can become an exchange item, and a waste exchange most likely exists for it.

Materials that are available for exchange are generated from a variety of sources, which include:

- By-products
- Damaged materials
- Expired products
- Obsolete and off-specification goods
- Overstock virgin products
- Surplus

Common materials that are available and wanted for exchange include categories such as:

- Acids
- Agricultural by-products
- Alkalis
- Ash and combustion by-products
- Chemicals
- Computers and electronics
- Construction and demolition debris
- Durables and furniture
- Glass
- Metals
- Miscellaneous
- Oils and waxes
- Paints and coatings
- Paper
- Plastics
- Refractory material
- Rubber
- Sand
- Services
- Shipping materials
- Solvents
- Textiles and leather
- Wood

Waste exchanges are used by a variety of organizations, including private sector waste generators, government agencies, solid waste district staff, recycling organizations, and material brokers. Materials exchange users can be anyone who handles surplus or unwanted materials, such as architects, administrative assistants, buyers, engineers, residents, consultants, custodians, environmental managers, government employees, procurement specialists, purchasing representatives, recycling brokers, shipping clerks, and storeroom managers.

Differences in the business models and processes for the waste commodity exchanges are also evident. Many of these exchanges serve as a meeting place for companies that would like to list materials and potential respondents, who then work out the details of payment, transportation, and storage themselves to facilitate exchanges. Some exchanges have an eBay-type Web posting system, whereas others produce printed periodicals. Some handle requests via the phone or fax; however, most utilize the Internet.

According to the EPA materials exchange, "Typically, the exchanges allow subscribers to post materials available or wanted on a Web page listing. Organizations interested in trading posted commodities then contact each other directly. As more and more individuals recognize the power of this unique tool, the number of Internet-accessible materials exchanges continues to grow, particularly in the area of national commodity-specific exchanges" [20].

Finally, the major difference among the exchanges dealt with the fee structures. Most exchanges are no cost, but some charge periodic membership fees or fees per transaction. Overall, waste exchanges have very minimal fees – just enough to cover the administrative costs. The American Plastics Exchange, for example, is the most expensive waste exchange, with a $360-per-year membership fee and $0 per exchange – still, a very cost-effective exchange. The typical per exchange fee, for those that did charge, was $5–$10 [20].

Success Stories

Waste exchanges have played an important role in assisting companies to identify and implement recycling and reuse opportunities. These efforts result in lower operating costs, reduced purchasing costs, reduced storage costs, enhanced corporate images, diminished demand for landfill space and incinerator capacity, and, ultimately, a cleaner environment.

"It is estimated that, by promoting the reuse and recycling of industrial materials through waste exchanges, the industry currently saves $27 million in raw material and disposal costs and the energy equivalent of more than 100,000 barrels of oil annually," as determined by the National Materials Exchange Network operating in Silver Spring, Maryland [21]. These savings often translate directly to the companies' bottom lines with stronger financial performance [21].

In 1998, its first year of operation, the Ohio Materials Exchange (Columbus, Ohio) exceeded initial expectations by exchanging over 2,600 US tons of waste and saving Ohio businesses $103,000 in disposable costs [22]. According to Dale Gallion, manager of quality assurance at Diamond Products (Elyria, Ohio), the company used to pay to have leftover metal powder scrapped [22]. After joining the waste exchange program, Diamond Products sold 8,000 pounds of metal powder for $14/pound. Additionally, the company accumulates about 1,000 pounds of metal powder each month and plans to continue using the exchange [22].

Another good example is the Massachusetts Materials Exchange [23]. "In the past 4 years, the Massachusetts Materials Exchange has moved over 2,000 US tons of materials, saving participants more than $100,000 in avoided disposal and purchasing costs" [23].

Waste exchanges are a cost-effective means of helping businesses save money, as well as helping to divert waste into usable raw materials. Advances in information technology over the past decade have served as a catalyst to promote the exchanges, and further allow exchanges to provide current information on both the materials available for use and the materials wanted, which helps business make better environmental and financial decisions [20].

These exchanges also play an important role in assisting waste generators in identifying and implementing recycling and reuse opportunities. For example, since the inception of the Ohio Materials Exchange, businesses using the service reported savings of over $13.5 million in disposal costs and diverted over 308,000 metric tons (340,000 US tons) from landfills [22]. These efforts result in lower operating costs and diminished demand for landfill space and burning capacity, which all lead to a cleaner environment. Natural resources are limited, and resources need to be conserved as much as possible [22]. The fewer raw materials used, the greater supply for future generations, and waste exchanges are helping achieve this goal.

Future Directions

The current trends within recycling collection and separation revolve around discovering more cost-effective methods and ways to enhance participation. Automated collection systems are gaining a great deal of popularity and will continue to proliferate as many organizations seek to reduce long-term costs.

Bibliography

Primary Literature

1. USEPA (2008) Municipal solid waste in the United States: 2007 facts and figures, EPA-350-R-08-101, Washington, DC
2. Fishbein B (1994) Germany, garbage, and the green dot: challenging the throwaway society. Inform Publishing, New York

3. Stessel R (1996) Recycling and resource recovery engineering: principles of waste processing. Springer, Berlin/Heidelberg/New York
4. Where it goes: aluminum recycling processes. SITA, UK, http://rethink.sita.co.uk/where-it-goes/aluminium/. Accessed 15 Nov 2010
5. Fix A (1997) Glass: reduce, reuse, recycle. Heinemann/Raintree, Chicago, IL
6. McKinney R (1994) Technology of paper recycling. Springer, Berlin
7. Goodship V (2008) Introduction to plastics recycling. Smithers Rapra Press, Akron, OH
8. USEPA (1988) Waste minimization assessment manual. Hazardous Waste Engineering Research Laboratory, Cincinnati
9. Petek J, Glavic P (1996) An integral approach to waste minimization in process industries. Resour Conserv Recycl 17: 169–188
10. Recycling collection systems, State of Connecticut, Department of Energy and Environmental Protection. http://www.ct.gov/dep/cwp/. Accessed 15 Nov 2010
11. Granger T (2007) Types of Curbside Recycling Programs. Earth911, http://earth911.com/recycling/curbside-recycling/. Accessed 31 Dec 2007
12. Curbside recycling, Earth911, http://earth911.com/recycling/curbside-recycling/. Accessed 20 Nov 2010
13. Wastes – resource conservation – conservation tools – pay-as-you-throw. The U.S. EPA, http://www.epa.gov/osw/conserve/tools/payt/. Accessed 20 Nov 2010
14. The U.S. EPA (2010) Wastes – resource conservation – common wastes & materials – paper recycling. http://www.epa.gov/osw/conserve/materials/paper/basics/grade.htm. Accessed 20 Nov 2010
15. The recycling process after collection what happens to materials when you recycle, The University of Oregon. http://pages.uoregon.edu/recycle/after_collection.html. Accessed 20 Nov 2010
16. Franchetti M (2009) Solid waste analysis and minimization. McGraw-Hill, New York
17. Franchetti M (2009) Case study: determination of the economic and operational feasibility of a material recovery facility for municipal recycling. Resour Conserv Recycl 52:535–543
18. (1994) A review of industrial waste exchanges. The U.S. EPA, solid waste and emergency response (5305), EPA-530-K-94-003
19. The U.S. EPA (2010) Recycling market development. www.epa.gov/epaoswer/non-hw/recycle/jtr/comm/exchange.htm. Accessed 20 Nov 2010
20. The U.S. EPA (2010) Waste and material exchange. www.epa.gov/epaoswer/non-hw/recycle/jtr/comm/exchange.htm. Accessed 20 Nov 2010
21. The National Material Exchange Network (2010) Waste exchange directory http://www.cftech.com/BrainBank/MANUFACTURING/WasteExchgs.html. Accessed 25 Nov 2010
22. The Ohio Material Exchange (2010) The Ohio material exchange: background. http://www.myomex.com/about.aspx. Accessed 25 Nov 2010
23. The Massachusetts Material Exchange (2010) Materials exchange: the benefits of recycling. http://www.materialsexchange.org/. Posted 13 Sept 2010

Books and Reviews

Avsar E, Demirer G (2008) Cleaner production opportunity assessment study in SEKA Balikesir pulp and paper mill. J Cleaner Prod 16(4):422–431

Beck RW (2004) Lycoming county material recovery facility evaluation. Pennsylvania Department of Environmental Protection, Final report

Berenyi EB (2002) 2001–2002 materials recycling and processing in the United States, 4th edn. Governmental Advisory Associates, Westport

Birgisdottir H (2006) Environmental assessment of solid waste systems and technologies: EASEWASTE. Waste Manage Res 24:3–15

Chang N, Wang SF (1995) The development of material recovery facilities in the United States: status and cost structure analysis. Resour Conserv Recycl 13:115–128

Covey SK (2000) Building partnerships – the Ohio materials exchange. Resour, Conserv Recycl 28(3):265–277

Covey SK (2000) Building partnerships – the Ohio materials exchange. Resources, Conserv Recycling. 28(3): 265–277

Davila E, Chang N (2005) Sustainable pattern analysis of a publicly owned material recovery facility in a fast-growing urban setting under uncertainty. J Environ Manage 75:337–351

De Alessi M, The dirty MRF. http://www.pacificresearch.org/pub/sab/. Accessed 20 Nov 2010

Diaz LF, Savage GM, Eggerth LL, Golueke CG (1993) Composting and recycling municipal solid waste. Lewis, Boca Raton

Dowie W, McCartney D, Tamm J (1998) A case study of an institutional solid waste environmental management system. J Environ Manage 53:137–146

enviro/05_enviroindex/25_toxics.html. Accessed 20 Nov 2010

EPA (1988) Waste minimization opportunity assessment manual. US Environmental Protection Agency, Cincinnati

Franchetti M (2008) One company's trash is another company's treasure: with the rising volumes of recyclable materials entering the waste stream, waste commodity exchanges have started receiving increased interest from companies looking trim their bottom line. Resource Recycling: N Am Recycling Compost J, Vol. 28, pp 40–42, February 2008

Franklin and Associates for the EPA (2000) Municipal solid waste generation, recycling, and disposal in the United States: 2000 facts and figures. June 2002

Haman W (2000) Total assessment audits (TAA) in Iowa. Resour Conserv Recycl 28:189–198

Harris E (2000) Cornell waste management institute update June 2000. Ithaca

Lui H (2003) Waste minimization at a nitrocellulose manufacturing facility. J Environ Studies 60:353–61

McCartney D (2003) Auditing non-hazardous wastes from golf course operations: moving from a waste to a sustainability framework. Resour Conserv Recycl 37:283–300

Mixed waste processing. NC division of pollution prevention and environmental assistance. http://www.p2pays.org/ref/01/00028.htm, Jan 1997. Accessed Feb 2006

Modern marvels: history of garbage (2004) Video produced by The History Channel

Ohio EPA (1997) Governor's pollution prevention award recipient: mahoning county's industrial waste minimization project. Fact Sheet #41

PEER Consultants and CalRecovery (1993) Material recovery facility design manual. CRC Press, Boca Raton

Petrell R, Duff S, Felder M (2001) A solid waste audit and directions at the University of British Columbia, Canada. Waste Manage Res 19:354–365

Rathje W, Murphy C (2001) Rubbish!: the archeology of garbage.

Rhyner CR, Schwartz LJ, Wenger RB, Kohrell MG (1995) Waste management and resource recovery. Lewis, Boca Raton

Ryding S (1994) International experiences of environmentally sound product development based on life cycle assessment. Swedish Waste Research Council, ARF Report 36, Stockholm, May 1994

Schianetz K, Kavanagh L, Lockington D (2007) Concepts and tools for comprehensive sustainability assessments for tourism destinations: a comprehensive review. J Sustain Tour 15:369–389

Shimberg SJ (1985) The hazardous and solid waste amendments of 1984: what congress did ... and why. The Environmental Forum. March 1985, pp 8–19

(2001) State solid waste management plan 2001. Ohio EPA – Division of Solid and Infectious Waste Management, Columbus

Strong DL (1997) Recycling in America, 2nd edn. ABC-CLIO, Santa Barbara

The Environmental Industries Associations Garbage then and now (2003) Washington, DC

The league of women voters (U. S.) Garbage primer (1993) Washington, DC

Tompkins JA, White JA, Bozer YA, Tanchoco JMA (2003) Facilities planning. Wiley, Hoboken

U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy (2001) Industrial Assessment Center Brochure 2001. Washington, DC

U.S. Environmental Protection Agency (1990) The nation's hazardous waste management program at a crossroads. Report no. EPA/530-SW-90-069. EPA, Washington, July 1990, 114 p

U.S. Environmental Protection Agency (1998) Household hazardous waste management: a manual for one-day community collection programs. Washington, DC

United States Congress of Technology Assessment (US Congress OTA) (1992) Green products by design: choices for a cleaner environment. OTA-E-541. U.S. Government Printing Office, Washington, DC

United States Environmental Protection Agency (2003) Municipal solid waste generation, recycling, and disposal in the United States: facts and figures for 2003. United States Environmental Protection Agency, Washington, DC

United States Environmental Protection Agency (2006) Municipal solid waste – basic facts. http://www.epa.gov/epaoswer/non-hw/muncpl/facts.htm. Accessed Feb 2006

United States Environmental Protection Agency (2006) Decision maker's guide to solid waste management, volume II. http://www.epa.gov/epaoswer/non-hw/muncpl/dmg2.htm, 1995. Accessed Feb 2006

United States Environmental Protection Agency (2006) Full cost accounting for municipal solid waste management. http://www.epa.gov/epaoswer/non-hw/muncpl/fullcost/docs/fca-hanb.pdf. Accessed March 2006

US Department of Energy, Office of Energy Efficiency and Renewable Energy (2001) Industrial assessment center brochure. Washington, DC

US Environmental Protection Agency (2006) Full cost accounting for municipal solid waste management. US EPA, Washington, DC

Wang Q, Dong L, Xi B, Zhou B, Huang Q (2006) The current situation of solid waste management in China. J Mater Cycles Waste Manage 8:63–69

Recycling Technologies

Giuseppe Bonifazi, Silvia Serranti
Dipartimento di Ingegneria Chimica Materiali
Ambiente Sapienza, Università di Roma Via
Eudossiana, Rome, Italy

Article Outline

Glossary
Definition of the Subject and Its Importance
Introduction
Recycling and Materials to Recycle
Future Directions: Innovative Control/Sorting Devices/Logics Integration in Recycling Plants
Bibliography

Glossary

Ceramic glass Transparent ceramic products with an appearance similar to that of glass. They are characterized by an amorphous phase and one or more crystalline phases.

Classification Set of mechanical actions carried out in dry or wet conditions, designed to perform a "classification" of particles systems according to their morphometrical (e.g., size-shape) attributes.

Comminution Set of mechanical actions carried out to reduce waste materials in particles of suitable size and shape to be properly handled and processed in order to liberate/remove contaminants.

Cullet Particulate solid product resulting from collection-comminution of waste glass.

De-inking Mechanical process that removes "ink particles" and "stickies" from waste paper.

Ferrous metal Magnetic metals mainly composed of iron.

Flotation Mechanical process that selectively separates hydrophobic from hydrophilic materials. Hydrophobic materials are forced to adhere to bubbles and float.

Fluff Fine fractions resulting from automotive shredder residue (ASR). *Fluff* is constituted by materials characterized by intrinsic low specific gravity (e.g., plastics, rubber, synthetic foams, textiles, etc.).

Municipal solid waste (MSW) All non-hazardous waste resulting from the collection of household, commercial, and institutional waste materials.

Non-ferrous metal Metals that contain no iron (e.g., aluminum, copper, brass, bronze, etc.)

Separation Set of mechanical actions carried out in dry or wet conditions, designed to perform a "separation" of particles systems according to their physical attributes (e.g., density, surface properties, electrostatic properties, magnetic properties, color, etc.)

Sorting Waste particle separation, usually carried out with optical-electronic recognition devices and logics.

Definition of the Subject and Its Importance

Recycling technologies can be defined as the whole of procedures designed to set up physical-chemical actions, at an industrial scale, that perform the recovery of materials and end-use products resulting from the collection of household or industrial wastes. The materials to be recovered and recycled, obviously, influence both processing technologies and plant layouts. In this section an in-depth analysis of the problems arising when suitable recycling technologies must be designed, implemented, and set up is presented with particular reference to paper, glass, metals, plastics, and textiles (not organics or **C***onstruction* and **D***emolition* (C&D) waste). Recycling technologies must be approached from a processing perspective, that is, by defining a sequence of steps and actions where the waste flow stream feed, and the different products resulting from the different sequential processing steps, are handled in order to produce one or more outputs of materials to reuse. Obviously, processing strategies and equipment must be selected with both low environmental impact and positive economic perspectives in mind. Dealing with waste often means dealing with complex products, that is, products constituted of one or more materials of interest but also of polluting material. The economic value, per unit of weight, of the materials to recover is usually low: recycling technologies thus must assure high production, while minimizing plant investments and management costs. From this perspective, a full characterization of the input waste streams and complete control of the different phases of the recycling process are a key issue when recycling technologies are selected and applied. In this section, for each of the different materials, methodologies, procedures, and logics are presented to preliminarily identify and quantitatively assess recycling technologies according to the characteristics of the materials to be recovered.

Introduction

The concept of recycling is an intrinsic part of nature. The different forms of life, at the end of their life cycle, decompose in the soil and become "compost," which helps plants grow. Furthermore, the organic materials resulting from decomposition of vegetation are food for bacteria and fungi. Bacteria and fungi are a food for earthworms, and earthworms are food for ants, beetles, and so on. This recycling chain can be framed in the theory of conservation of mass, originally stated by Heraclitus (530–470 BCE) [1]:

- πάντα χωρεῖκαὶ οὐδὲν μένει: everything changes and nothing remains still

and after by Nasir ad-Din Tusi (1201–1274) [2]:

- a body of matter cannot disappear completely. It only changes its form, condition, composition, color and other properties and turns into a different complex or elementary matter and finally clearly outlined and formalized by M. Lomonosov (1711–1765) and A. Lavoisier (1743–1794) [3]:

- the mass of a closed system (in the sense of a completely isolated system) will remain constant over time.

Recycling can thus be considered as something related to nature, and as consequence, to humans. Since the beginning of time people have needed to find a way to dispose of and/or to recycle waste. Obviously, technology influenced and continues to influence recycling strategies. In early pre-industrial times waste was mainly constituted of combustion residues, wood, bones, bodies, and vegetable waste. A simple recycling approach, mimicking what happens in nature, was to dispose of these wastes in the ground. Wastes thus became compost, helping to improve soil. Nearly 4,000 years ago there was a recovery and reuse system of bronze scrap in Europe. Composting is known to have been a part of life in China (2000 BCE). Ancient rubbish dumps excavated in archeological digs reveal only tiny amounts of ash, broken tools, and pottery. In Knossos (Crete, Greece), traces of landfill sites exist, dating from 3000 BCE, where waste was placed in large pits and covered with earth at various levels. The evolution of humans from nomadic hunter-gatherers to farmers increased waste production. Waste could no longer be left behind, and it soon became a growing problem.

The First World War, the Great Depression, and the Second World War each contributed to recycling. Recycling, in fact, was a necessity for many people to survive or for a nation to support the war effort. Nylon, rubber, and many metals began to be recycled during this period. The practice of recycling continued in many countries after the Second World War came to an end, especially those nations with a high dependence on resources, such as Japan. However, in other countries, with the post-war years' economic boom, consciousness about the importance of applying policies addressed to recycling and recovering materials and products at the end of their life cycle rapidly decreased. In the 1960s and 1970s, the importance of recycling returned with the Environmental Movement (e.g., the first celebration of Earth Day was in 1970), and a constant and continuous growth followed.

The mass production of the Industrial Age is the main cause of the past low level of consciousness about the importance of recycling: when products can be produced, and/or purchased, at low cost, at least if compared with the income, it often seems more economical to just throw away old items and purchase new ones. This "disposable goods" mentality and the corresponding problems of waste disposal have created environmental problems that today all countries face today when adopting recycling technologies.

What are the benefits related to a wider use of recycling technologies?

The first benefit is obviously related to the reduction of the environmental impact of human activities. The possibility of significantly decreasing the industrial use of non-renewable resources, such as primary raw materials and fossil fuels, by utilizing products resulting from recycling represents an important step forward toward environmental protection and energy savings; both aspects strongly contribute to a reduced exploitation of natural resources. The second benefit is the reduction of the environmental impact related to waste dumping. Less waste disposed of means less natural sites to select and manage for waste storage and less risk to the environment in terms of soil contamination and surface water and groundwater pollution. The third benefit is related to better design and manufacturing of products, from the simplest (e.g., paper, glass, metals, or plastic containers) to more complex ones (e.g., household appliances, cars, etc.) from the perspective of recycling (e.g., ease in dismantling) at the end of their life cycle.

What problems are to be faced in increasing the use of recycling technologies both quantitatively and qualitatively?

The lack of the public acceptance toward recycling and the subsequent low growth of the related **W***aste*-**D***erived*-**P***roduct***s** (WDPs) market are the main factors negatively affecting recycling in quantitative terms [4]. Such problems have progressively decreased thanks to (1) the definition of a clear traceability route of the WDPs, (2) the technical-economical improvement of recycling products, and (3) new legislation at the national and/or international level stimulating recycling and utilization of recycled products. In qualitative terms, an obstacle to wider "up-to-date" utilization of recycling technologies is related to concerns about the application of innovative technologies inside existing recycling layouts. Today, a great amount of equipment (e.g., comminution, classification, and separation units), and related operative technologies, is easily available on the market. Sometimes, lack of knowledge of specific control tools, necessary for correct handling and control of the equipment, can affect final recycling plant layouts and overall quality of the plant itself in terms of capacity to adapt to new possible future market requirements. From this perspective, the utilization of equipment and inspection tools for waste products quality control are fundamental for a modern, efficient, and profitable recycling technologies implementation. Both equipment and control device systems must be selected and fully integrated according to: (1) waste feed attributes, (2) possible feed variation, and (3) required final products characteristics. These three aspects must always be taken into account in the definition and analysis of the recycling technologies utilized for all the materials described in this section.

For paper, glass, metals, plastic, and textiles, the above-mentioned aspects play a different role and assume a different importance. In paper recycling feed characteristics are relatively easy to control. Final product characteristics, that is, the quality of the recovered fibers, is preeminent. In glass recycling, the main problem is feed quality, especially for the recycling of glass collected from **M***unicipal* **S***olid* **W***aste* (MSW), where the presence of "ceramic glass" can strongly impact the further processing and final quality of recovered glass fragments. With metal recycling, the aim is to maximize the "correct" recovery of non-ferrous metal alloys. The achievement of this goal strongly influences the development of innovative sorting/detection logic in order to assure the requested final products' characteristics. In plastic recycling, both feed and final products influence the selection of the separation devices and strategies, as well as the sensing technologies required for quality assessment during the different processing stages. Finally, dealing with fiber recycling requires maximizing source material identification and its preliminary separation. Process and technologies are quite different for textiles and carpets; carpets are much more difficult to recycle than textiles. Here, we have briefly outlined the different problems that must be taken into account in the recycling and recovery of the most common waste materials. In the following, such procedures will be analyzed and the related processing/control devices and actions illustrated.

Recycling and Materials to Recycle

Recycling Technologies: Paper

Paper is usually made from raw material wood pulp and fiber. Vegetable fibers are mixed and "cooked" until the fibers are sufficiently softened, chemicals (e.g., lye) are added to enhance and accelerate softening. The pulp is then "screened" over a screening media. Water is dropped off and/or evaporated. The material is then pressed for further water removal in order to obtain the "paper sheet." The quality and arrangement of the fibers affects the overall quality of the final manufactured paper material [5]. With this in mind, recycling technologies applied to waste paper are primarily aimed at maximizing recovery of the fibers.

Paper production dates back to the ancient Egyptians (e.g., papyrus paper). Around 200 BCE Cai Lun, a Chinese court official, made paper from tree bark and old fish netting. Its production was considered as a remarkable secret and only 500 years later were the Japanese able to acquire the secret. Papermaking spread to the West when some Chinese paper makers were captured by Arabs after the defeat of the Tang troops in the Battle of Talas River (751 AD). The first European paper mill was built at Jativa (Valencia, Spain) around 1150. From that time to the fifteenth century, paper mills were located mainly in Italy, France, Germany, and England; by the end of sixteenth century they were located all over Europe. In 1719, Rene de Reaumur, a French scientist, observed wasps chewing slivers of wood and building their nest starting from such a fiber paste. The use of wood fibers for papermaking started from this observation.

One of the key points in papermaking is an appropriate handling of connected fibers. They can come from a number of sources including cloth rags and cellulose fibers from plants and trees. The use of cloth in the papermaking has always produced high-quality paper. The presence of cotton and linen fibers in the mix creates papers for special uses. From this perspective, cotton and linen rags can be profitably re-utilized for fine-grade papermaking (e.g., bank notes, certificates, letterhead, resume paper, etc.). Rags are usually cuttings and wastes from textile and garment mills. Also, the paper itself can be profitably recycled. The constituting fibers can be reused five to seven times before they become too short and brittle. Many paper-based-products can be manufactured from recycled waste paper: fruit trays, corrugated cardboard, egg cartons, ceiling tiles, plasterboard, sound insulation panels, etc. Waste paper collection and recycling produces a number of positive environmental effects:

- Less timber is used for wood pulp production, which has a positive impact on biodiversity, that is preservation of valuable wildlife habitats and ecosystems, such as old-growth forests that are not replaced by managed plantations, very often constituted by allochthonous species, usually fast-growing conifers
- Waste disposal reduction
- Energy and water savings, as there is no need for pulping to turn wood into paper. Such a savings depends on paper grade, processing, mill operation, and proximity to a waste paper source and markets
- Considerable reduction of emissions into the air and water (no bleached is usually required in recycled paper)
- Lower greenhouse gas production; the larger the amount of waste paper re-used, the lower the emissions will be

The main problems faced in waste paper recycling are as follows:

- Collection criteria addressed to simplify waste paper handling and further processing
- Identification of polluting elements and suitable processing strategies in place to remove them
- Quantity and the quality of pollutants (e.g., effluents) discharged to water

Waste Paper Characteristics Waste paper mainly originates from pre-sorting and collection of consumer waste and/or industrial waste. The following attention will be primarily addresses consumer waste, which is mainly constituted by:

- Newspapers, magazines, telephone directories, and pamphlets
- Cardboard
- Mixed or colored paper
- White office paper
- Computer printout paper

Waste paper is usually subjected to sorting according to its origin and characteristics. Such characteristics have been quantified at in Europe with the definition of the European List of Standard Grades of Recovered Paper and Board [6]. Waste graded papers are then pressed and handled as bales. Bales can be thus assumed as the secondary raw materials fed to waste paper recycling plants.

Different waste paper processing strategies will be thus adopted according to the presence of contaminants, to collected paper grade, and to their final re-use. Waste papers contaminants are usually constituted by:

- Materials and/or products not directly utilized in paper manufacturing, such as metals (e.g., nuts, screws, foil, cans), plastics (e.g., films, bags, envelopes), cloth, yard waste, leather, and dirt;
- Materials and/or products directly utilized in paper manufacturing, such as:
  - Inks and toners.
  - “Stickies” (e.g., adhesives, coatings, pitch, resins, etc.): these tend to deposit inside paper manufacturing equipments (e.g., wires, press felts, dryer fabrics, calendar rolls, etc.) causing problems, mainly machine down-time. Furthermore, they are difficult to remove due to their neutral density and resulting particles flow characteristics.
  - Coatings: these are usually constituted by inorganic fillers (e.g., $CaCO_3$, TiO_2, clay, etc.) and polymeric binders. Fillers have to be removed from the pulp and lower the overall yield of the recycling process. The presence of wax-treated papers (e.g., cardboard) negatively affects recycled paper quality in terms of weak and slippery properties. Furthermore, wax tends to deposit on equipment.
  - Fillers (e.g., $CaCO_3$, TiO_2, clay, etc.): their removal is compulsory when specific paper product, as tissue paper, have to be produced.
  - Papermaking additives (e.g., starch, gums, dyes, etc.): among the most difficult to handle are dyes. Their incorrect removal can affect recycled

paper new coloring. In some cases, wet strength additives can prejudice the further waste paper re-use.

As a result, waste paper characterized by high quality grades (e.g., paper-mill production scrap and office waste) requires simpler processing and can be profitably applied as a primary paper pulp substitute in applications such as paper printing and tissues. Waste paper of intermediate grades (e.g., newspapers) must be subjected to a stronger processing, mainly for de-inking, and can be used again by the newspaper industry. Finally, waste paper of lower grade is utilized for packaging and board.

Waste Paper Recycling Technologies In waste paper recycling the selection and the sequence of the processing units is strongly influenced, as previously outlined, by the characteristics of the waste paper (e.g., grade) and the presence and typologies of contaminants [7]. The latter influence the process, not only by their composition, but also by their chemical-physical attributes (e.g., size, shape, density, surface properties, solubility, strength, etc.). In the following, the main units, related actions, and/or potential problems are reported and described with reference to waste paper processing:

- Re-pulping, "to pulp" waste paper
- Screening, for fine contaminant removal
- Cleaning, for contaminant removal
- De-inking
- Water and solid waste treatment

Re-pulping The re-pulping operation is the first and one of the most important processing stages in paper recycling. Correct re-pulping is a pre-requirement for efficient downstream operations (e.g., cleaning, screening, flotation, etc.). Incorrect re-pulping can damage fibers, preventing their "correct" re-use [8]. A re-pulper is thus a device that converts recovered paper into a slurry of well-separated fibers and other waste paper components by performing mechanical, chemical, and thermal actions [9]. In order to fulfill this goal, a re-pulper has to satisfy specific conditions, that is:

- Contaminant detachment from fibers, without performing comminution (the larger the contaminant, the easier its removal)
- Correct mixing between waste paper, H_2O, and chemicals to liberate fibers, limiting at the same time their cutting
- Contribution to large debris removal

Re-pulping can be carried out in batch or continuous conditions. In batch conditions waste paper, H_2O, and chemicals are all charged at the beginning of the process and are removed, all at once, at the end of the process, then the process starts again. In continuous conditions waste paper, H_2O, and chemicals are continuously added to the pulper, as the pulped product is continuously removed.

Screening Screening is usually performed by forcing the pulp to pass through a sieve. The sieving surface is characterized by holes and slots of different sizes and shapes. The main goal of this phase is removing contaminants (e.g., bits of plastic, glue globs, etc.) and at the same time to realize a first separation of "short" fibers. Screening performances are influenced by many variables, the most important being:

- Feed pulp characteristics: fiber size and shape, quantity and quality of debris
- Screening device characteristics and operative conditions: screen surface (e.g., flat or cylindrical), screen hole size and shape, screen surface cleaning mechanism, fed pulp flow rate, solid-water ratio, stock temperature, etc.

Cleaning Cleaning is mainly applied to remove heavy contaminants. Separation is usually based on centrifugal forces, frequently using hydrocyclones. These devices are constituted by a cone-shaped (e.g., tapered) cylinder. Pulp is fed under pressure to the device, the rotational movement produces, inside the hydrocyclone, two vortexes create the separation: heavier particles are thus recovered in the bottom part and the lighter ones in the upper part. During this stage, metals, inks, sand, and dirt, are usually removed. The cleaning stage is influenced by many variables, including:

- Fiber and contaminant characteristics (e.g., size, shape, density, and quantity);
- Selected device architecture and setting: cylinder/cone size, inlet and/or outlet geometrical

characteristics, vortex finder diameter and length, cylindrical section height, cone angle, feeding pressure, pulp dilution, etc.

Deinking Pulp deinking removes printing ink and "stickies" (sticky materials like glue residue and adhesives). Deinking is usually performed in two steps: (1) washing and (2) flotation. Small particles of ink are thus preliminarily rinsed from the pulp with water by washing, then large particles and "stickies" are removed with the help of chemicals and air bubbles by flotation.

Froth flotation technology has been developed and used for many decades in the mineral processing industry before the technology was adopted by the pulp and paper industry for the deinking of waste papers at the beginning of the 1960s [10]. A flotation process is based on the surface properties of particulate solids systems when suspended in a fluid. Particles according to their, natural or caused, hydrophobic or hydrophilic characteristic tend to adhere to bubbles and float. During flotation deinking, pulp is thus fed to one, or to a bank of, flotation cells, where air (e.g., bubbles) and chemicals (e.g., surfactants) are also present. The surfactants cause flotation of the ink and sticky materials. Air bubbles carry the ink particles to the top of the cell/s, where the foam is continuously removed, realizing the required pulp deinking.

The most significant differences between deinking flotation and mineral flotation are specifically linked to the particular characteristics of waste paper pulp suspensions [10], that is:

- The large size class distribution and shape of the particles to float (e.g., ink particles), as well as their surface properties. Ink particles, in fact, can vary from about 1 μm–1 mm, they are generally hydrophobic, except for water-based inks. Large particles are usually flat shaped, and other techniques, as previously outlined, such as screening, with slots down to 0.1 mm and centrifugal cleaning, are also used to remove the various impurities of waste paper pulp suspensions (e.g., pressure sensitive adhesives, hot melt glues, plastic films, etc.)
- The low density of the particles to be removed from the deinking pulp: polymeric particles with specific gravity close to that of the water. Mineral particles (e.g., fillers, kaolin, and $CaCO_3$, utilized for paper coating), in the size range of about 1 mm, should generally not be removed
- The presence of flocs or networks (e.g., cellulose fibers typically of 1–3 mm in length and 10–30 μm in diameter according to wood essences originally utilized) that tend to flocculate up to constitute about 1% of the volume, in the separation zone of deinking cells, as the turbulence level is decreased
- The need to add chemicals to the re-pulped waste papers, both to realize a better release of the ink particles from the fibers, at the same time enhancing the flotation process (e.g., calcium soap and caustic soda or other deinking chemicals to be used under alkaline or neutral conditions) and the various chemicals introduced with the waste papers (e.g., surfactants used in the coating color)

After flotation, if necessary, pulp is further beaten, or "refined," in order to separate, as much as possible, fibers, avoiding fibers bundles. When white recycled paper is required, pulp is bleached with hydrogen peroxide or chlorine dioxide.

Water and Solid Waste Treatment Production of both virgin and recycled paper gives rise to pollutants that are discharged to water (e.g., effluents). Studies providing comprehensive comparative evaluation of the environmental impact linked to the effluents generated from recycling plants and those from paper mills demonstrated the environmental impact of the former is lower than that of the latter. In any case, environmental problems related to paper waste recycling are, with reference to the other recycling technologies, further described in this chapter, those presenting a higher impact [7]. The different waste paper processing stages, and related utilized technologies, are, in fact, always carried out in wet conditions and with a large quantities of water and chemicals.

Water. Four key parameters have to be fully monitored in the waste water resulting from waste paper processing: total suspended solids (TSS), biological oxygen demand (BOD), chemical oxygen demand (COD), and chlorinated organic compounds (AOX). De-inking is the main cause of TSS and BOD, and sometimes these parameters are comparable with the same produced processing virgin pulp. On the other hand, COD and AOX are always lower in effluents resulting from waste paper processing. Waste water must be properly processed before it can be re-utilized or before release in the environment. The significant decrease, in recent years, of Cu, Cr, Pb, Ni, and Cd in printing inks dramatically contributed to reduce heavy metal presence in waste water, sludge, and final recycled-paper-based-products.

Solid wastes. The sludge resulting from waste paper processing contains a solid fraction ranging between 30% and 50%. It is mainly constituted by short fibers, fillers, and ink from the de-inking process. Their relative proportion depends on waste paper source characteristics and processing strategies applied to obtain a final product of the required characteristics. Usually the wastes are sent to dumps. In recent years, different attempts have been made to further process and/or re-use them: composting [11], and removal of clay [12] and other fillers [7] for re-use or their utilization for energy production [13].

Emissions to air. Direct emissions from the process of making recycled paper itself are minimal and considered to be relatively insignificant, although little research has been done in this field. Gaseous and particulate emissions to air are produced when the thermal utilization of sludge generated by the pulp and papermaking processes is carried out. Combustion presents many advantages [13], including reduction of the disposed solid mass and volume leading to lower disposal costs, destruction or reduction of the organic matter present in the sludge, and energy recovery. Critical points related to the adoption of this solution are:

- The **L***ow* **H***eating* **V***alue* (LHV) characterizing wet sludge, requiring preliminary dewatering and/or drying treatments to bring solids content above 30–35% in order to enable a self-sustained combustion and
- The presence of potentially hazardous elements (e.g., sulfur, chlorine, cadmium, and fluorine), that requires a complete gas cleaning

At the end of all the above-described processing stages recycled pulp fiber finally enters the machine for manufacturing recycled paper sheets. Waste-paper-recovered-fibers can be used alone, or blended with virgin ones to achieve better strength, or smoothness, of the final paper product.

Recycling Technologies: Glass

Glass is made of three relatively simple raw materials, silica sand, limestone, and sodium carbonate, which are melted together at high temperatures (about 1,500°C). Additives can be included to modify some properties, such as color, refractive index, durability, etc. [14].

Examples of glass manufactured goods can be found from several thousand years BCE, when such material was used for ornaments. In the Renaissance period, glass use increased. Vessels, bottles, and other glass containers started to be produced and utilized for both decorative and everyday use. At that time glass manufactured goods were expensive to produce. Large-scale production started with the Industrial Revolution and mass production of glass containers began at the onset of the twentieth century. Together with the increase in production and larger use came the problem of handling glass waste. Glass manufacturers produce a large quantity of products of different characteristics that are addressed to different uses. Glass's physical properties, at high temperature, are close to that of a viscous fluid, and as a consequence it can be worked, by craftsmen or on an industrial scale, to obtain final products of practically nearly infinite number of shapes and characteristics. For this reason glass can be found in, according to its composition and use, several products, ranging from those commonly used at home (e.g., bottles, vases, jars, mirrors, etc.), to those utilized in the automotive sectors (e.g., windscreen) and in industry (e.g., fiberglass for the production of *G*lass *R*einforced *P*lastics (GRP), *G*lass *R*einforced *C*ement (GRC), special thermal and/or acoustic insulating panels, X-Ray and cathode tubes, etc.). It is thus easy to understand that waste glass production, and its recycling to produce mainly "new" glass containers, assumes great importance.

Glass is one of the materials that is most often recycled. It presents a series of positive characteristics: it is non-absorbent and does not confer flavors and odors; it resists high temperatures, such as those required for cleaning after its use; its strength and mechanical resistance are indispensable for multiple fillings and reuse. These characteristics make glass containers suitable to be used over and over again. Selective waste glass collection and recycling provide great benefits:

- Reduction of environmental impact related to its disposal

- Conservation of the non- renewable raw materials (quartz sands) required for its production
- Energy savings
- Reduction in the quantity of solid urban waste

The problems to face in glass recycling can be summarized as follows:

- The definition of collecting criteria able to simplify the further processing
- The identification of polluting materials and the set up of suitable processing strategies to remove them
- The separation of "broken glasses" (*cullet*) according to their color

Recycled glass mainly comes from the selective collection of solid urban waste (bottles, jars, various containers, etc.), usually done by citizens, and only partially from the refuse of glass goods manufacturing and/or glass-based products at the end of their life-cycle. As a consequence, waste glass collection represents one of the most critical steps of the entire recycling process, and the following recycling technologies and separation strategies are strongly conditioned by the criteria and the methods followed during collection. The quality of the collected materials can be quite different, according to the level of knowledge and, more generally, the "education" of the people involved. As a matter of fact, the quality of the glass collected for recycling can strongly differ from region to region or, in the same city, from district to district.

The following discussion is based on urban waste collection as the source of the glass. It is important to consider the final destination of the recycled glass, which can be identified by the classical market categories where glass in commonly utilized, that is: (1) container production, (2) construction industry, (3) special concrete production (e.g., partial substitution of aggregates by glasses), (4) road pavement (e.g., special asphalt where the coarse fraction is partially substituted by glass), (5) abrasive products, (6) wool glass, etc.

The recycling technologies described for glass recycling will be primarily addressed to producing an economically valuable *cullet* to use to make new containers. Recycled glass is not equally re-utilized in all the above-mentioned market sectors. Only a small fraction is, in fact, re-utilized in fiberglass, bricks, concrete, and asphalt production. This is mainly for two reasons: (1) *cullet* quality sometimes does not fit well with the quality standards required in some of the these sectors and (2) the glass container industry is the most interested in waste glass reuse (due to the high cost of primary raw materials versus the relatively lost cost of each single glass container).

Cullet characteristics must satisfy strict conditions to be re-utilized for container production. These characteristics are primarily related to both presence of polluting elements and color of the fragments. Furthermore cullet size class distribution is another important parameter to control. Usually particles around 1 or 2 cm are preferred both for handling and quality control purposes.

Contaminant removal and *cullet* color sorting are the main goals when recycling glass. Furthermore, such goals must be reached using a process that does not produce too fine particles. As a consequence recycling technologies must be designed to fulfill these goals.

***Cullet* Contaminants Definition** There are two classes of contaminants: materials not constituted by glass (e.g., ceramics, stones, masonry, organics, and heavy metals) and glass fragments of the wrong color, that is *cullets* whose color characteristics are different from that of the class they belong to (cross-contamination).

Non-glass materials. Ceramics and stones, which have melting points higher than that of the glass, remain un-melted inside the vitrified matter and as a consequence, even if present in a small amounts, degrade the mechanical characteristics (resistance) of the manufactured products (bottles, jars, etc.). Furthermore, they can seriously damage glass processing equipment, increasing maintenance costs. Lead and heavy metals, according to their high volume weight, settling on the bottom of the fusion crucibles and have a corrosive effect on the refractory material, causing, in some conditions, the perforation of the refractory material itself. Optical sorting devices are commonly used to identify and automatically remove *non-glass materials*. Among polluting materials special attention has been addressed, in recent years, to ceramic glass. This material rapidly increased its presence in waste glass products, mainly due to the introduction on the market of a large amount of ceramic glass manufactured goods, such as dishware, cookware, etc. [15]. Such material, even if seems quite similar to classic glass, is characterized by a different behavior (i.e., higher fusion point) when melted inside glass furnaces, where cullets are usually fed together with natural raw materials (quartz sands) [16]. As a consequence, the presence of ceramic glass reduces the production rates of the furnace, which needs to be shut down to be cleaned more frequently, and sometimes causes damage that requires the furnace to be rebuilt or replaced. Classical optical sorting devices are practically "blind" to ceramic glass, as its physical-chemical characteristics are similar to those of glasses.

Cullet cross-contamination by color. Glass has, according to its color, a different destination of use and, as a consequence, different market value. The use affects the value of the glass containers; as a consequence, white glasses have higher value than the so-called half-white or colored glass (brown, yellow, green). *Cullet*s, that are collected without distinction of color can be primarily used for the production of green glass and only in part for the production of yellow glass. The production of white glass requires that only *cullet* of that color be employed. Cross-contamination can thus represent a problem because it always contributes to depreciate the *cullet's* value. For this reason *cullet* optical sorting by color is extensively utilized.

***Cullet* Recycling Technologies** On the basis of the previously mentioned washing goals, specific processing layouts have to be defined. Each of them will be constituted by one or more units, in series and/or in parallel, and each one specialized to perform a specific action. In the following the main units, and related actions, are reported and described with reference to waste glass fragments (*cullet*) processing:

- Hoppers and conveyor belts, to perform *cullet* handling
- Manual sorting of macro-contaminants (e.g., ceramics, metals, plastics, stones, etc.)
- Crushing and screening, to reduce glass fragments in two or three size fractions, avoiding the production of fines, for a better handling of the glass waste, and to perform, in some cases, also a separation of contaminants
- Ferrous and non-ferrous metals separation, respectively, by electro-magnets and eddy-current separators, for metals and non-metals not manually sorted
- Light contaminant (e.g., paper and plastic) removal
- *Cullet* fine contaminants and color sorting by magnetic and/or optical devices, to define different *cullet* quality classes according to market specifications

Hoppers and Conveyor Belts Storage-feeding and conveying equipment constitutes the skeleton of the plant providing glass waste transportation for the action performed. Conveyors thus assume an important role because they have to operate continuously to assure a constant feed to the different units of the recycling plant. *Cullets* are strongly abrasive, so conveying, and related mechanical devices, are subject to strong abrasive actions. Furthermore, *cullet* handling produces fine particles and dusts. These aspects, if not fully controlled (e.g., sealed bearings, enclosed gear boxes and other moving parts protection, suction units for dusts collection) can produce severe damage to conveying units, as well as severe environmental working conditions in the plant. Table 1 are details some of the characteristics that a conveyor belt has to satisfy to be correctly utilized inside a glass recycling plant.

Manual Sorting of Macro-Contaminants Human-based sorting continues to be commonly applied on the coarser fraction. Manual sorting allows removing the larger pieces of contaminants, easily detectable and removable by trained personnel. Today, this phase is often performed by automatic sorting equipment, utilizing different principles, mainly optical, for contaminant detection. Among macro-contaminants, in recent years, ceramic glass has become a problem. As its machine-based identification is difficult, manual sorting continues to be utilized to perform its removal.

Crushing Waste glass comminution, usually performed by crushing equipment, is carried out to reduce the collected waste glass materials to particles of suitable size and shape in order to be properly handled and processed to remove contaminants and to separate glass into suitable classes of colors to produce new containers or other products. After crushing a sieving stage is usually carried out. *Cullet*s are thus "sorted" according to their size.

Recycling Technologies. Table 1 Conveyor belt design principles and/or maintenance good practices with reference to waste glasses handling

Cleats	Conveyors without cleats are preferred, because they can represent an obstacle to conveyor surface cleaning by scrapper bar. Their presence is negative if moisture is present.
Bearings	Bearings and other revolving parts (e.g., pulleys) have to be regularly cleaned or, better, should be sealed.
Drivers	Drivers should be selected in order to assure flexibility according to end markets variability.
Surface	Belt constituting materials have to be selected taking into account cullets' abrasive characteristics and related wear effects. Rubber belt are the most commonly utilized but wearing effects are severe, for this reason vibratory conveyors are more and more utilized. They are less subject to wearing but can represent a further source of pollution (e.g., presence of fine metallic parts in the final cullet concentrate).
Moisture and/or organic residues	Moisture can severely affect *cullet* handling, especially with reference to sorting operations.

Glass, different from aggregates, when subjected to comminution (e.g., crushing) tends to produce, for its chemical-physical characteristics, a lot of fines (pieces smaller than 1–2 mm); such a behavior increases if comminution equipment is not properly selected, that is, if the comminution unit and its operative conditions do not take into account glass mechanical characteristics. A comminution unit usually develops its action through the application of four forces on the materials: impact, shear, abrasion, and compression. The quantitative relationships among them, in terms of cause-effect, strongly differ according to the equipment. For glass comminution, impact forces are those to be primarily applied, allowing a low production of fines. The production of fines causes several problems: (1) loss of product (e.g., recycled glass fragments below 2–3 mm are usually rejected from the market), (2) a double loss of money (e.g., unsold material and rejected fines have to be disposed of), and (3) higher costs for comminution equipment maintenance. Glass fines are extremely abrasive; they can enter the gears and bearings of traditional comminution equipment and cause serious damage. Furthermore, the presence of fines produces the same mechanical problems in the other equipment constituting the processing layout. When selecting a comminution unit it is thus important to evaluate not only the capacity, but also the degree of flexibility characterizing the crusher, that is, the ability to vary power and speed. These two parameters can be modified according to changes in required crushed glass characteristics in respect of the initial feed and possible requirements of the end market. Table 2 describes some of the impact crushing equipment most commonly used to perform waste glass comminution.

Screening and Air Classification Screening and air classification are usually performed for two purposes: (1) particle size control and (2) contamination removal. The proper use of screening and air classification is to cut *cullets* into particles belonging to different size class ranges. Such a division can be performed both to facilitate *cullets*' further processing and to obtain final marketable products characterized by a population of specific and well-defined particle size. Screens and air classifiers can be profitably utilized as "separators" when polluting elements (e.g., wood, paper, iron, steel, aluminum, plastics, etc.) are characterized by a size and/or a shape different from the cullet population to classify. In this case, the setting of a "threshold," for the geometrical attributes, allows separating pollutants from glass. Furthermore, when air classifiers are used, the different densities of the polluting materials, compared to those of glass, represent an important factor in performing an improper "air classifier based separation." In both cases, moisture is a restricting factor for both classification and separation efficiency.

In the selection of screening and/or air classification devices some important factors have to be taken into account: (1) particle size distribution of the feed, (2) particle size distribution after the different classification stages, and (3) market tolerance, that is,

Recycling Technologies. Table 2 Characteristics of the comminution units mainly utilized to perform waste glass crushing

Comminution units	Characteristics
Hammer Mills (HM)	Movable hammers mounted on a rotating shaft hit and/or throw glass against mill chamber or other glass. As a result comminution is realized. Cullets are recycled inside the hammer until they do not reach a size lower than the aperture of a grid installed at the exit of the mill chamber. Comminution is efficient, that is, a relatively small size glass grain can be achieved in just one comminution stage. The wear rate of the hammer is high.
Rotating Drums (RD)	A spinning rotor with bars attached to the outside is responsible for the comminution. Glass particles breakage is mainly due to: (1) the impact of the rotating bars on glass and (2) the glass projection against "special plates" installed on the inside the chamber. Crushed material is finally discharged when particles size is lower than the space, adjustable, between the rotating bar and the plates. The wear rate is lower in comparison to hammer mills.
Vertical Shaft Impactor (VSI)	Crushing is performed by a revolving vertical rotor. The material is fed from the top. The mechanical actions resulting for the interaction of the revolving rotor with the glass and with the chamber walls produce the requested glass size reduction. A screening system, in a closed loop with the impactor, allows obtaining the desired *cullet* size. VSI allows obtaining large productions, but a pre-broken feedstock, usually obtained utilizing as primary crusher an HM, is required. The wear rate is minimized; the resulting *cullets* are round shaped.
Impact Crusher (IC)	IC utilizes a continuous breaker bar which is mounted horizontally in the rotor. Glass is fed and blown against *adjustable* aprons. Replaceable liners are installed inside unit. The broken product passes through an open discharge. As for VSI, IC allows to efficiently handle a large volume of throughput of product below 10 mm. *Cullets*, according to IC comminution actions, are characterized by an angular or sub-angular shape.

precision required with respect to end-user tolerance versus possible presence of under-/over-size in the final classified product. Table 3 describes some of the screening and air classification devices commonly utilized in the waste glass recycling sector.

Ferrous and Non-ferrous Metal Separation Ferrous (e.g., metals containing iron) and non-ferrous metals (e.g., aluminum, copper, lead, zinc, tin) are among the most common contaminants in waste glass. Their origin can be related to the presence of caps, lids, special labels, bottle neck wrap, etc. The presence of these materials even at low level can produce severe damage to the production process (e.g., deposits and chemical reactions in the furnaces, presence of hot spots, clogging or jamming of the injection devices, etc.) and can compromise the final product (container) mechanical characteristics and quality.

Ferrous contaminants are easily removed by magnetic separation. Usually magnetic separation is performed in two stages. A first stage, at the beginning of the recycling process, in order to remove ferrous contaminant before the further processing stages (comminution, classification, sorting by color, etc.) and a later stage, where usually it is applied to remove the finer particles of metals not separated during the previous processing stages. The selection of a magnetic separator is strongly influenced by: (1) waste feed rate, (2) amount of ferrous contaminants, (3) *cullet* stream depth, (4) contaminants size, and (5) value of the magnetic field to apply to perform separation. Non-ferrous metal contaminants are not affected by the presence of a magnetic field; for this reason eddy current separators are usually utilized for their collection [17] [18]. Eddy current separation devices and related separation strategies are strongly influenced by non-ferrous metal particles size, shape, and conductivity. In this latter case, different from magnetic separation, particles' attributes play a preeminent role. Table 4 details the equipment commonly utilized to perform ferrous and non-ferrous metal separation.

Light Contaminants Removal Light contaminants such as plastic, paper, wood, corks, etc. do not

Recycling Technologies. Table 3 Characteristics of the classification units mainly utilized in the waste glass recycling sector

Classification units	Characteristics
Trommels (TRO)	A TRO is a rotating cylinder whose surface is characterized by apertures of a specific size and shape where *cullets* can pass through according to their size. For its characteristics TRO can play both as a classifier and a separation unit. Contaminants as plastic bottles or bags are not able to pass through trommel surface apertures and can thus be easily removed. A trommel can be designed to fit practically any size glass processing operation. In some cases hot air is blown to facilitate waste glass feed drying, helping the screening and the following processing stages.
Vibratory Screens (VS)	VS usually work in a closed loop with crushing unit. As for TRO *cullets* can pass through, or not, the sieving surface, according to their size, usually larger particles are fed again (closed loop) to the comminution unit (crusher) or rejected if recognized as pollutants. According to the requested *cullets* size class characteristics, vibratory screens can allow to classify particles population up to – 75 μm. When such dimensional grades are requested obviously throughputs are lower. Usually multi-deck VS, up to five, are utilized when the final *cullet* product has to be divided in several dimensional classes.
Air Classifiers (AC)	These types of classifiers are based on the utilization of air flow to classify particles. Different from sieving, particles are thus classified/separated according to their size, shape, and density. An AC is a vessel where an air flow is generated. Wastes are usually fed from the top, coarse (larger or heavier) and fine (smaller or light) particles follow a different path according to flow characteristics, that can be set to achieve the required classification/separation. Classification/separation can be achieved simply utilizing gravity or/and free or forced vortex generated by static vanes or a dynamic classifier wheel. The equilibrium between the forces determines the "cut point." ACs represent a good alternative to screening, especially when the materials to classify are characterized by the presence of large fraction of fines (particles below 250 μm). When a well classified coarse fraction is requested, air classifiers are utilized prior to screening.

constitute, at least in principle, a problem in glass recycling. They burn and volatize at the temperatures of the glass furnaces. Today, new applications requiring granular products are sensitive even to the presence of small amounts of "organics"; for this reason the recycling processes have to take into account the removal of these kinds of contaminants. Such a goal is easily reached by simple screening or by air-screening, that is creating a sort of fluidized bed where heavier particles, *cullet*, remain on the site and lighter or finer particles can be easily removed by air flow.

Fine Contaminants and Color Originally *cullet* optical sorting was performed following an optical-based analogical approach. It was developed and implemented to perform control-separation actions:

- To remove all contaminants not removed in the previous physical processing stages (e.g., venting, classification, eddy current separation, magnetic separation, etc.)
- To perform a separation of *cullet* by color

Normally, ceramics, stones, and opaque particles are sorted before color sorting is applied. In recent years, the increasing presence, inside waste glass, of ceramic glass, which is almost impossible to separate via classical recycling technologies, has resulted in research related to optical sorting to solve this problem with a new class of devices based on X-rays and/or H*yper*S*pectral* I*maging* (HSI) technology.

Sorting was originally performed manually. Such an approach is labor-intensive. Being based on human senses it is strongly affected by human sorter experience, by the size of the materials to sort, and by time: the level of attention of the worker is strongly affected by time. With technological developments, *cullet* sorting was, and in some cases continues to be, adopted as detection unit, laser beam technology based devices, and scan line cameras. Table 5 describes the devices, and related architectures, commonly utilized to perform *cullets* sorting.

Sorting device characteristics. These are related to the optic detector/s size and arrangement. The final

Recycling Technologies. Table 4 Characteristics of the separation units commonly utilized to perform ferrous and non ferrous contaminants separation in the waste glass recycling sector

Separation units	Characteristics
Overhead/Cross Belt Magnets (O/CBM)	The magnet is usually installed above the glass waste flow stream (e.g., conveyor belt). The overhead magnetic field has a belt moving across its surface at approximately a 90° angle to the materials flow. Metals particles are thus attracted, removed from cullets and discharged as the moving belt of the separator turns away from the magnetic field the metals particles. Sometimes, especially when O/CBM strategy is applied to control the possible presence of ferrous particles in the final product, "simple" high intensity magnets are utilized. Collected particles are thus discontinuously removed.
Magnetic Head Pulleys (MHP)	MHP is usually installed at the end of a conveyor belt, beneath the belt. Ferrous particles are thus held to the belt while cullets can be discharged. With the decrease of the magnetic field ferrous particles leave the belt and can be properly recovered.
Magnetic Drums (MD)	MD commonly installed inside feeder chutes, between chutes and conveyors. Their behavior is similar to MHP. Glass waste stream passes over the magnetic drum, ferrous metals are held by the drum, as non-magnetic materials continue their flow. Ferrous materials hold to the drum until a divider provides to its discharge.
Eddy Currents Separators (ECS)	ECS is realized by spinning a magnetic rotor with alternating polarity at high speed. The magnetic drum of the ECS induces electric currents (eddy currents) within the volume of each particle flowing in the proximity of the drum. The effect is that non-ferrous metals passing over the drum are subjected to an ejecting force that throws away non-ferrous metals pollutants from the waste glass stream. The main difference from ordinary magnetic separation is that the magnetization of particles is not induced by the alignment of the internal magnetic domain with the external field. The efficiency of the eddy current separation process is highly dependent on the size of the feed particles.

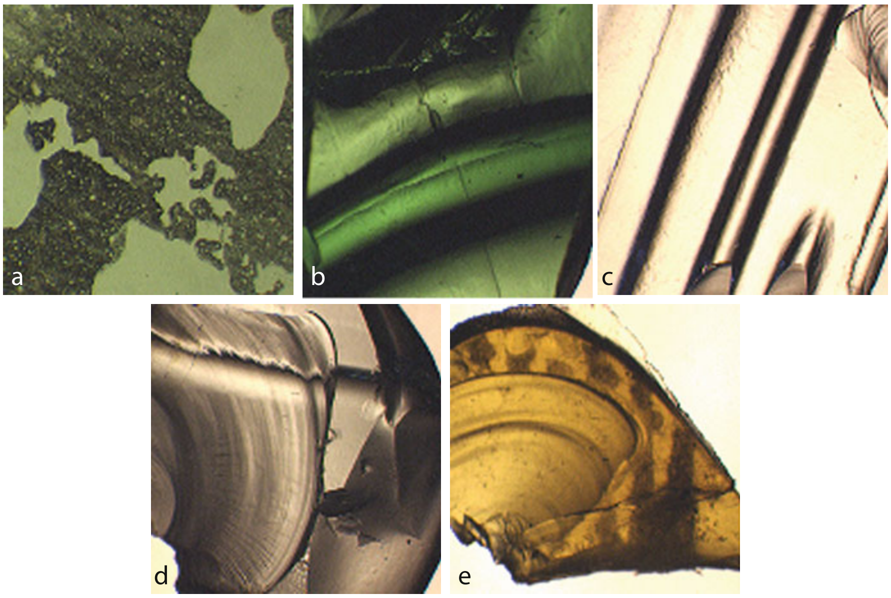
sensing architectures, in fact, are influenced by optical information acquisition and its further handling: both factors dramatically influence the sorting architectures, usually based on a pneumatic blast, enabling the modification of the *cullet's*, and/or polluting particle's, trajectory, after recognition. Furthermore, flow characteristics can influence the selection. To realize an optical sorting (Table 5), the flow has to be, at least in principle, constituted by particles forming a mono-layer (e.g., *cullet* fed to the color-sorting unit by a vibrating conveyor belt, which keeps the glass in a thin layer). In these conditions glass fragments can be analyzed by the laser beams. Recent equipments are so fast that they can test, according to its dimension, the same cullet several times. As a consequence the larger the glass fragment, the better its control can be carried out. The influence of the "anomalies" can be thus reduced when each cullet is analyzed by more than one detector and more than one time. This is not possible with smaller pieces: they can pass, for their dimension and for the flowing conditions, unsorted, or diffraction/refraction effects (e.g., presence of marked cleavage or surface anomalies) can be so strong that detectors are practically unable to analyze them. As a consequence, such a technique cannot be profitably used with the entire size range of cullets. Materials of smaller dimensions, less than 2 mm, resulting from the processing-cleaning stages, not being correctly investigated, are usually disposed of, with the resulting financial loss and negative environmental impact. For this reason, in recent years, a new class of sorting devices based on imaging has been developed and implemented (Table 5). Imaging allows breaking down the investigation limits, with detection resolution linked to array detector resolution and optics magnification. Obviously the geometrical constraints, linked to the presence of the pneumatic devices, for particle removal remain; however, the imaging approach constitutes a big step forward allowing: (1) a better detection and (2) the potential to perform *on-line* certification of products.

Recycling Technologies. Table 5 Characteristics of the detectors, and related logics, commonly utilized to perform *cullet* sorting

Sorting units	Characteristics
Laser Beam Technology Based (LBTB)	Detection is based on the evaluation of the "characteristics" of the energy and the spectra received by a detector after the *cullets*, and/or polluting particles, were crossed by a suitable laser beam light. Such an approach presents two technological limits, related to: (1) the constructive characteristics of the equipment and (2) the material characteristics. The sorting logic is mainly analogical. An *on-off* logic is applied.
Image Analyzers (IA)	IA allow to perform sorting on the basis of the *cullets'* detected colors. CCD (**C***harged* **C***oupled* **D***evice*) scan line cameras are utilized. They present the advantage, in comparison with LBTB, that practically do not have any detection limitation in terms of geometrical resolution, being the investigated scan line dimension is the only function of the lens. Furthermore, colors are better detected.

Material characteristics versus their optical recognition. Cullet attributes, surface status (e.g., dirty or clean) and characteristics (e.g., fragments of bottle neck or jar with or without thread, bottle or jar bottom, etc.) can strongly interfere with the measurements. Measurements are, in fact, based on the evaluation of the transmitted energy received by a detector after the *cullets* are crossed by an energizing source (e.g., standard or laser beam light). In these measures surface characteristics and the status of each cullet play a decisive role in the further response of the detector (Fig. 1).

Ceramic glass recognition. As previously outlined, the frontier in *cullet* optical sorting is represented by ceramic glass recognition. The only two strategies extensively used and designed to reduce the presence of ceramic glass contaminants are source reduction and manual sorting. The problem with reduction at the source is that usually citizens, in spite of public education campaigns, confuse transparent-glass-like contaminants with normal glass, invalidating both curbside and door-by-door collection. As a consequence, some steps of the sorting process are still carried out manually by "trained personnel" who try to recognize ceramic glass fragments prior to crushing, looking at their shape or evaluating their reflective characteristics. Such an approach is expensive, unreliable, and represents a real and important problem for the whole glass recycling sector. Infrared sensors are sometimes used to detect opaque cullet within the "dirty" recyclable glass in order to separate ceramic/ceramic glass and stone fragments from glass cullet. No entirely effective and low-cost solution has been found to date for ceramic glass *on-line* automatic sorting in recycling plants. X-ray sorting techniques have been recently proposed as a solution for ceramic glass identification [19]. Anyway, it must be considered that the use of X-ray equipment in a plant requires appropriate shielding and must follow strict rules to protect workers from exposure, with an increase in costs and environmental and safety problems. Glass sorting, originally based on analog devices, utilizing laser beam technology, has moved towards digital image techniques [20, 21]. Scan line color cameras are thus seeing greater use in this sector to implement selection strategies addressed to identify opaque objects inside the flow stream and/or to separate cullets according to their color [22]. Technology in this field, although sophisticated, remains practically "blind" with regard to the identification of ceramic glass materials. Spectrophotometers should be able, at least in principle, to identify these contaminants; however, they are usually only able to work on a point-by-point basis and are not able to cope with real-time sampling/sorting architectures such as those required in glass recycling plants [23]. They are used, in several industrial fields, mainly at the laboratory scale. A new class of sensing devices based on HSI has recently created new potential for the *on-line* recognition of glass and ceramic glass fragments inside glass recycling plants [24]. A more detailed description of such a technique is reported in the later section Future Directions.



Recycling Technologies. Figure 1
The spectral response of the cullets, suitably energized, is based on the evaluation of the transmitted energy received by the detectors. The detected energy is thus influenced by the status (dirty or clean) and the characteristics (fragments of bottle neck or vase with or without thread, bottle or vase bottom etc.) of the cullet surface. (**a**) 16 × 16 mm image field of dirty *light green cullet*. (**b**) 16 × 16 mm image field of *dark green cullets* (*bottle bottom*). (**c**) 2 × 2 mm image field of *white cullets* (*threaded bottle neck*). (**d**) 2 × 2 mm image field of *half-white cullets* (*bottle bottom*) and (**e**) 2 × 2 mm image field of *brown cullets* (*bottle bottom*)

Recycling Technologies: Metals

Human history and progress are linked to the discovery and utilization of metals [25]. Thanks to these materials, humans have been able to interact and modify the environment, performing advances in agriculture, warfare, transport, etc. The Industrial Revolution, from steam to electricity, was conditioned by metals. Even what, and how, people eat was and is strongly influenced by metals.

The first use of metals dates back to about 7000 BCE in Anatolia (Turkey), where some Neolithic communities started to replace handmade stone knives and sickles with “hammering” native copper. The tools worked as well as their stone equivalents and lasted far longer. The first examples of extractive activities belongs to 4000 BCE. Deep shafts were cut into the hillside at Rudna Glava, in the Balkans, to excavate copper ore. Mining was considered a sort of ritual activity; as thank for the exploited metals, fine pots, bearing produce from the daylight world, were placed in the mines as a form of recompense to propitiate the spirits of the dark interior of the Earth [26]. Thousands of years later, humans started again to understand the importance of preserving the environment and, together with more stringent economic reasons, started to apply more metal recycling.

Metals present many advantages; they can easily be recycled because a specific material can melted several times without losing its properties. Metals commonly utilized in the recycling sectors mainly derive: (1) from the collection and processing of post-consumer metal products and (2) from metal industrial wastes (e.g., working residues, metallic scraps, etc.). Recycling strategies are strongly affected by the previously mentioned origins of the metals. Furthermore, metals’ value is strongly affected by several costs, that is, the quality of recovered products (e.g., composition, contaminants residues, etc.), (3) recycled product market, and

(4) metal market value. Together with these costs, their processing (e.g., collection, transport, sorting, etc.) and waste disposal settlements also play an important role. Metals to recycle can be divided in two different families, that is: ferrous and non-ferrous metals. Such a distinction is important because it dramatically influences recycling strategies and related adopted technologies. Ferrous metals scraps are mainly constituted of iron and steel scraps, primarily obtained from automotive dismantling and household appliances (e.g., large kitchen appliances, washers and dryers, etc.). Such waste is usually collected and preliminarily sorted in different classes of products of different grade before being sent to a recycling plant or directly to metal refiners. Wastes resulting from their processing (e.g., wood, plastics, fibers, etc.) are, according to their quantity and physical characteristics, totally or partially recovered; the unrecovered fraction is sent to a landfill. Another lesser source of ferrous metals results from the processing of the bottom ashes produced by incinerators.

Non-ferrous metals scraps are mainly constituted by aluminum, copper, zinc, and lead. Aluminum is the main scrap deriving from household waste (e.g., cans, containers, etc.); the others primarily result from waste from industrial and commercial activities.

Ferrous Metals Iron and steel are the main materials utilized in many industrial sectors: building and construction, automotive, chemical, operative equipments, etc. These materials are so common in use for several reasons: (1) relatively low costs, (2) high availability, (3) good mechanical attributes, (4) easiness in working, and (5) because they can also, at least in principle, be easily recycled, the main reason being linked to the ease in recovering them due to their magnetic properties.

Iron and steel are obtained from raw materials (e.g., iron ores) and/or 2) from recycling. Different production methods are thus utilized: **B***last* **F***urnace* (BF) and **B***asic* **O***xygen* **F***urnace* (BOF), when primary raw materials are utilized, and **E***lectric* **A***rc* **F***urnace* (EAF) when recycled products are employed. Metal scrap recycling allows reduction of both energy-production costs (e.g., less energy is required for produced unit of weight when scraps are re-melted instead of using iron ore) and environmental impact (e.g., reduced exploitation of primary raw materials such as iron ores, limestone, and coal necessary when primary metals are produced by BF or BOF). Furthermore, a corresponding decrease of CO_2 emission is achieved, with a further environmental benefit.

Metal recycling is a well-established practice and it will continue to grow with the increased availability of automotive-derived scraps. In the future, however, this source will probably be reduced as more motor vehicles are designed with plastics and/or polymeric-based composites. In a quantitative way, steel represents the most recycled product, more than aluminum, paper, and glass together and greater than all other metals combined (e.g., aluminum, copper, nickel, chromium, zinc).

Metal Scraps' Sources and Characteristics Metal scraps can be obtained from many sources:

- *Home scrap*, that is, the scrap (e.g., working production waste, defective parts, etc.) derived from a manufacturing process. In this case, waste is directly re-melted. There are no problems related to metal scrap quality, as the material is constituted only of the metal to recycle
- *Industrial scrap* usually consists of the wastes produced in iron- and/or steel-manufacturing plants, mainly leftover product resulting from specific manufacturing actions. Such wastes are usually sold and re-used in foundries
- *Post-consumer scrap* is metal waste derived from products that have reached the end of their lifecycle (e.g., industrial equipment, cars, metals structures, home appliances, etc.). In this class of scrap, contamination, mainly the presence of non-ferrous metals (e.g., aluminum, copper, zinc, and lead), can represent, in some circumstances, an important problem to face and solve

Other iron and steel intermediary products, and/or waste, play an important role as recycled products: steel-making slag and flue dust (resulting from BF, BOF, and EAF), waste sludge and filter cake (resulting from BF and BOF), spent pickle liquor, and mill scale.

One of the most important steps in the development and set-up of metals recycling systems is creating appropriate strategies to identify and sort metals into groups presenting similar characteristics. Such a "grouping" must be carried out according to specific rules defined by steel-makers and market requirements.

Ferrous Metals Recycling Technologies On the basis of what has been previously outlined, specific processing layouts have to be defined. Each is constituted by one or more units, in series and/or in parallel, and each is specialized to perform a specific action. The main actions, and related units, are reported and described below with reference to ferrous metal wastes processing:

- Manual sorting (e.g., non-metal miscellaneous material detachment, large non-ferrous metal separation) and preparation (e.g., cutting of large metal manufactured goods, removal and dismantling of specific metal unit, etc.)
- Crushing and screening, to reduce metal scraps to easy-to-handle pieces to be directly fed into furnaces or subjected to further classification-separation actions
- Separation of the different metal fractions into groups characterized by similar composition attributes
- Testing and/or sorting of the different resulting ferrous metals end-life-goods-derived-material, to characterize and certify different classes of products according to market requirements

Manual Preparation and Sorting Metal manufactured goods to recycle are usually constituted of units of large dimensions (e.g., automobiles, metal structures, etc.); for this reason they must first be reduced to easy-to-handle pieces, both for further processing and/or for direct re-use inside furnaces. For these reasons, shears, hand-held cutting torches, crushers, or shredders are commonly utilized. After this preliminary stage, manual sorting, if required according to ferrous metal waste to recycle, is carried out. At this early stage of the process large contaminants that are easily detectable with human senses (e.g., car batteries, plastics, foams, wood, non-metallic elements, etc.) are removed.

Materials Handling Conveying units are mainly usually conveyor belts. For the characteristics of the feed, especially at the early stages of ferrous metals handling (e.g., relatively large pieces of materials), handling equipment is of rough construction. Ferrous wastes are usually stockpiled and the primary feeding, after the preliminary operation outlined in the previous paragraph, is realized by cranes. Ferrous metals are abrasive; as a consequence, all the different parts of the equipment utilized for conveyors are subject to strong abrasive actions; also, the presence of fine particles and dusts must be carefully checked and reduced to avoid mechanical problems and to assure good environmental working conditions.

Comminution After manual preparation and sorting, further size reduction actions are applied to scraps. Comminution actions are different according to destination of the materials, that is: (1) direct feeding to the furnaces or (2) further processing. In the first case, large scrap materials are milled utilizing shears, flatteners, and torch-cutting and turning crushers. The resulting pieces are then compacted by baling or briquetting in order to increase the apparent density of the scrap aggregates to re-melt, trying to avoid their possible floating in the mold. In the second case, crushing actions are usually applied, the goal being to reduce scraps to suitable dimensions to allow their processing (further crushing stages, sieving, separation, sorting, etc.).

A comminution unit usually develops its action through the application on the materials of four forces: impact, shear, abrasion, and compression. The quantitative relationships among them, in terms of cause-effect, strongly differ according to the equipment. For ferrous metals comminution, impact, and shear forces are those to be primarily applied, in order to optimize metal scraps and minimize fine particle production. Coarser fractions are mainly constituted by iron and steel, finer fractions usually contain the residues. Finer fractions can be divided in heavy and light (*fluff*) fractions. They present different characteristics in composition according to constituting particles weight. **A***utomotive* **S***hredder* **R***esidue* (ASR) can be considered as the main source of metals. ASR heavy fraction mainly contains aluminum, stainless steel, copper, zinc, and lead. ASR *fluff*, for quantities and characteristics, represents an important class of ferrous metals end-life-goods-derived material of particular interest for the recycling sector. *Fluff* represents about 25% of the weight of a car. It is usually constituted by materials characterized by intrinsic low specific gravity (e.g., plastics, rubber, synthetic foams, textiles, etc.). When processed to perform their recovery, they pollute the materials presenting higher specific gravity (i.e., copper, aluminum, brass, iron, etc.), constituting parts of the electrical devices of the vehicle that, for their shape, size (e.g., wires, metal straps, slip rings, wipers, etc.) and utilization remain concentrated in the lighter products. Such "polluting agents," for their intrinsic characteristics, are not well removed by classical separation techniques. The development and application of efficient washing strategies for *fluff* could dramatically reduce waste and environmental pollution, allowing, at the same time, an increase in energy recovery through pure sorted polymer re-use. Furthermore, the potential to use finer *fluff* fractions to produce energy could contribute to increasing the full recovery of such kinds of products. To reach this goal the quantity and the quality of the metal contaminants have to be strongly controlled in order to not prejudice the quality of the final fluff-based fuel. Always with reference to car dismantling, ASR heavy fractions contain large quantities of both ferrous and non-ferrous metals. Their recovery is usually performed adopting recycling technologies based on heavy media and eddy current separation.

Shredding is usually the main comminution action applied. Processing layouts, embedding this phase, are usually applied to automobile hulks and to the so-called white goods, that is, stoves, refrigerators, washing machines, etc. The most utilized class of shredders are those based on the use of **S***wing*-**H***ammer* **S***hredders* (SHS), subordinately, **R***otating* **D***rums* (RD), are also employed. The main characteristics of both types of equipments are described in Table 6.

Recycling Technologies. Table 6 Characteristics of the comminution units mainly utilized to perform ferrous metals shredding

Comminution units	Characteristics
Swing-Hammer *Shredders* (SHS)	The input material is fed from the side. Fed material flow is controlled according to the energy required for comminution, usually greater of one order of magnitude in respect of minerals. Material is transported into the relatively narrow gap between the impacting tools and the lower part of housing, where it is subject to an intense deformation and comminution. The material, which has become sufficiently small, is discharged from the chamber of comminution by means of grates. The configuration of the discharge grates can vary. Normally grate is placed above the rotor and in some cases a second one below it. According to different comminution chamber size and shape, rotating of the rotor in respect of the feed, grates position and configuration, different comminution actions can be thus performed. In the last decade a lot of efforts have been addressed to investigate to utilize SHS with vertical mounted rotors. Such comminution units are actually mainly utilized in the processing of metallic cuttings, waste wood, paper waste, etc. Their possible full and systematic use in ferrous metal recycling sector could embed several advantages, that is: (1) a lower residence time of the metal particles inside the comminution chamber (e.g., higher flow rate and less energy consumption), (2) a better liberation and (3) a lower compaction of the liberated thin-walled metal pieces. A more systematic utilization of SHS, with vertical mounted rotors, is strongly linked to a further development of shredders' mechanical architectures and utilized materials characteristics.
R*otating Drums* (RD)	A spinning rotor with chains or bars attached to the outside is responsible for the comminution. Ferrous metals fragmentation is mainly due to: (1) the impact of the rotating chains or bars on metals and (2) metals projection against "special plates" installed on the inside the chamber. Crushed material is finally discharged when particles size is lower than the space, that can be properly set, between the "rotating unit" and the wall. This equipment is commonly utilized as the primary crusher. For their characteristics do not allow the degree of flexibility in terms of operative conditions as those allowed by SHS.

Separation Separation technologies are applied when the shredded materials to recover are composed of different families of particles characterized by different physical-chemical attributes and different relative composition, texture, and shape. For these reasons automatic and in-series handling-separation strategies are required, as simple manual sorting or separation are unable to recover in an efficient and economically profitable way the different materials. Table 7 lists the equipment commonly utilized to perform separation of end-life-goods-derived products; those resulting from car dismantling represent the main source of complex ferrous metal waste to recover. Because they are constituted of different particles of different magnetic properties, specific weight, color, chemical composition, etc., they require different separation strategies [27–29].

Testing Different from other recycling-derived materials, the possibility of performing a rapid test on samples collected from the different recycled ferrous metal flow streams is particularly important, especially with reference to alloys. Alloys of similar grades and composition are usually difficult to discriminate. Specific attributes are thus evaluated both by expert personnel and by specific tests. Recognition, via

Recycling Technologies. Table 7 Characteristics of the separation units commonly utilized to perform separation in the ferrous metal recycling sector, with particular reference to ASR

Separation units	Characteristics
Belt Magnets (BM) *and Drum Magnets* (DM)	BM and DM are usually utilized in the first stage of processing, which is on coarser fractions as they result from primary crushing. Permanent and/or electromagnets are usually utilized. Separation is realized adopting BM and DM. In the first case the magnet is located between pulleys around which a continuous belt travels. In the second case the magnet is installed inside the rotating shell, metals particles are attracted, removed from the other non magnetic fractions and discharged as the moving belt of the separator turns away from the magnetic field the metals particles. Following this approach iron and steel cannot be separated from nickel and magnetic stainless steels. An improper separation can negatively influence the further melting stage. For this reason, hand sorting, to reduce the contamination of the ferrous products, is usually performed after this stage.
Eddy Currents Separators (ECS)	ECS is realized passing the waste products to separate into magnetic field, as a result, eddy current induced in the non-ferrous metals, produces ejecting forces that throw away non-ferrous metals the waste feed flow. ECS is commonly applied after the first magnetic separation stage (e.g., DM or DM). The most utilized ECS architecture is based on an inclined ramp. The material is thus fed to the ramp. The ramp surface is usually constituted by stainless steel. Under the ramp surface a series of magnets is positioned. Due to the eddy current non-ferrous metals are deflected sideways. Separation is realized according to the trajectory followed by the different classes of materials. Other separation architectures are based on the use of a rotating cylinder or a conveyor belt: magnets are positioned around the rotating axis of the cylinder or fitted inside the head pulley, respectively. As in the previous case separation is realized according to materials trajectory variations.
Heavy Media Separators (HMS)	HMS are based on the utilization of a medium constituted by a finely milled solid (e.g., magnetite or ferrosilicon) and water. According to the solid/water ratio the density of the medium can vary. Usually such a value is between the value of the specific gravity of the two classes of materials to separate, so that a sink and a float product is obtained. In this process the recovered materials are then washed and dried. The fine particles of the heavy media are recovered, by magnetic separation, from the slurry resulting from product washing and re-utilized inside the process. Decreasing the size of the particles to separate, also separation efficiency decreases for the increasing effect of viscosity, in this case cycloning is utilized.

human senses and analytical equipment, is thus performed to evaluate characteristics such as color (e.g., copper and brass distinction), apparent density and hardness (e.g., lead distinction from copper and brass), magnetic properties (e.g., iron and stainless steel), presence and attributes of spark patterns as they results from abrasion test, chemical reaction to reagents, chemical and X-ray spectrographic analysis (e.g., alloys composition), thermal behavior (e.g., melting point), etc. All the above-mentioned approaches are expensive and are difficult to implement *on-line*.

Sorting As outlined in a previous paragraph, sorting is important because it allows removal of contaminants from the different ferrous metal flow streams handled in the recycling plant. Different sorting strategies addressed to recognize different metal scrap constituents have been developed and used. Table 8 lists the devices, and related architectures, commonly used to perform metal scrap sorting.

Non-ferrous Metals Aluminum is the main non-ferrous metals being recovered and recycled. The main source of aluminum scraps comes from the packaging, transport and homeware industries. Aluminum can be "infinitively" recycled. Its re-melting achieves several environmental goals: considerable energy savings (about 95%) in comparison with the energy required to produce it from bauxite ore, a consequent reduction of primary non-renewable raw materials exploitation, and a reduction of overall emissions (airborne dusts and CO_2). The automotive sector significantly increased, in recent years, its use of aluminum; for this reason scraps from end-of-life vehicles represent the main source of aluminum for recovery

Recycling Technologies. Table 8 Characteristics of the sorting units commonly utilized to perform sorting in the ferrous metal recycling sector, with particular reference to ASR

Detection units utilized for sorting	Characteristics
Portable Optical Emission Spectrometers (POES)	POES can be utilized to perform the on-site sorting and identification of metals. Such an approach, even if not reaching the precision of the corresponding laboratory device, is quite useful to perform fast quality control, usually well satisfying recovered products grade requirements. POES is able to detect up to 90% of the currently produced grades of steel.
Image Analyzers (IA)	IA allow to perform sorting on the basis of the detected color. IA belongs to the first class of devices utilized for metal scraps sorting. Adopting this approach, zinc, copper, brass, and stainless steel are commonly well sorted. Even with technological improvements, both in terms of speed of processing (the same pieces can be checked several times), sensor quality (better discrimination in terms of recognizable colors), and resolution (minimum identifiable scrap piece), IA is not efficient when slightly different alloys have to be recognized.
Laser-Induced Breakdown Spectroscopy (LIBS)	The detection architecture is based on the analysis of the optical spectrum, or fingerprint, of a small spot on each metal particle that is evaporated using powerful laser pulses. Even though very powerful, the techniques show some limitations. The most important thing is that it is particularly sensitive to the status of the scrap surface. Presence can negatively influence measurements because laser pulses can penetrate for just few Å in the surface.
X-ray Fluorescence Spectroscopy (XRF)	XRF is based on the emission of X-rays emission inside an XRF unit. The fluorescence radiation generated by the atoms when they release the energy, after the excitation stage is collected and analyzed. Both emitted wavelengths, and the energy released are functions of the elements constituting the waste sample. Correlating the emission intensity it is thus possible to evaluate the content of a specific element within the sample.

and reuse. Other non-ferrous metals commonly recycled are copper, zinc, and lead, but as outlined in the previous paragraph the recycled quantities are not comparable with aluminum and they are also relatively easy to recover. For this reason, the following recycling technologies will be described with reference to automotive aluminum scraps.

Aluminum from Automotive Scraps Aluminum scrap is usually classified into two categories: cast and wrought aluminum alloys. Recycling of aluminum scrap introduces several technical problems to the secondary ingot market. Because of its high reactivity, aluminum cannot be refined pyrometallurgically, as with copper or iron scrap. Therefore, the aluminum scrap can be recycled only by blending and dilution in order to obtain a specific alloy. Wrought aluminum alloys contain low percentages of alloying elements (e.g., silicon, magnesium, copper, and zinc), less than about 4% of the total. Casting aluminum alloys contain the same elements as wrought, but in greater amounts (the silicon content in cast alloys can range up to 22%).

Actual scrap sorting technologies produce mixtures of cast and wrought aluminum not suitable for recycling in wrought alloy production.

The wrought fraction of these mixtures has a higher value; if selectively collected its reuse as wrought aluminum alloys would prevent unnecessary downgrading. Moreover, aluminum alloys, used in vehicle manufacturing, are increasing and the recovery of aluminum alloys is quite interesting. Considerable efforts have been devoted to substitution of steel and cast iron with lighter materials such as aluminum alloys and polymers. Steel replacement by aluminum alloys could have a counterbalancing effect in the case of success in selection of cast and wrought aluminum.

Non-ferrous Metals Recycling Technologies The same processing steps and strategies previously described for ferrous metals can be also applied for non-ferrous metals scraps. Such phases, for scrap originating from end-of-life vehicles (one of the main non-ferrous metals sources, in particular aluminum) can be synthetically identified in:

- Collection, dismantling (e.g., removal of re-usable vehicle parts as engines, doors, glass, seats, etc. and hazardous parts as batteries, fluids, etc.) and/or manual sorting (e.g., metal miscellaneous material detachment, large ferrous metal separation)
- Crushing and screening, to reduce non-ferrous metal scraps in pieces easy to handle for further processing
- Separation of the different non-ferrous materials in classes of products characterized by composition attributes
- Testing and/or sorting

Collection, Dismantling, and/or Manual Sorting Non-ferrous metal manufactured goods to recycle can be constituted by pieces of different dimensions, ranging from aluminum cans up to a "jet airliner," thus, according to their size, they can be directly handled or must be cut to be properly handled. Often, non-ferrous metal parts are linked (e.g., bolted, welded, etc.) with other materials and have to be liberated. Hand dismantling and/or sorting in some cases is necessary, but in many cases it is time consuming and inefficient for the dimension and the degree of locking of the different constituting materials. For these reasons comminution/separation actions must be applied adopting specific processing layouts.

Materials Handling Conveying units are mainly conveyor belts of rough construction. Non-ferrous wastes are usually stockpiled and the primary feeding, after collection, dismantling, and/or manual sorting is done by cranes. Ferrous metals are abrasive, and as a consequence all the parts of the equipment used for conveyors are subjected to strong abrasive actions. The presence of fine particles and dusts also has to be carefully monitored and reduced to avoid mechanical problems and to assure good environmental working conditions. Furthermore, problems related to explosive characteristics of aluminum dust have to be taken into account.

Comminution and Screening Comminution and screening are currently applied to reduce metal scrap to different size classes that are sent for proper processing and recovery of the different materials constituting the feed. The considerations developed with reference to ASR metal fraction comminution, in terms of equipment and characteristics (Table 9), can also be directly applied to ASR non-ferrous metals.

Recycling Technologies. Table 9 Characteristics of the comminution units mainly utilized to perform non-ferrous metals size reduction

Comminution units	Characteristics
Alligator (AS) and/or *Guillotine Shears* (GS)	AS mimics, to perform cutting, the behavior of an alligator mouth. The device is constituted by two jaws, one fixed and the other mobile. The main advantage in the use of AS is its versatility in terms of different aluminum scraps it is (e.g., ship, aircraft, automotive vehicles and in general large objects) able to cut. Furthermore better detachment is allowed of different metals constituting a specific piece to dismantle. When a GS is utilized, aluminum scrap is placed underneath a cutting blade, which drops down onto the scrap creating the cut. GS are characterized by higher power ratings and productivity than AS.
Impact Shredders (IS)	IS as *hammer mills* are among the most used devices for aluminum scrap size reduction. Fragmentation is realized for two co-occurring effects: (1) hammers impact being the main one, and (2) scraps projection the secondary one, against mill chamber internal surface. *Impact crusher*s are often also utilized. They use the same milling actions but invert the relative effects.
Rotary Shredders (RS)	RS utilized, as main milling actions: (1) cutting and (2) impact for their architecture are normally utilized to process light metal scraps (e.g., foil and beverage cans and containers).

Separation Aluminum-based material separation technologies are mainly based on magnetic [27], eddy current [17], air [28], and sink-float separation [29]. The main characteristics of the different separation units, based on the previously mentioned physical principles, are reported in Table 10.

Among the different separation approaches, sink-float separation (S&FS) plays an important role when non-ferrous metals products have to be separated and/or refined before the final re-melting stages. Such products, in fact, are usually composed of a wide range of materials characterized by different density, shape, and size class distribution, strongly affecting sink-float separation. For this reason separation is performed in different stages. After a preliminary removal of *non-magnetic-fine fractions*, a two- or three-stage density separation is carried out. With a typical three-stage density separation, low-density plastics, foam, and wood are usually preliminary removed (*cut-off* density: 1 g/cm^3), then, utilizing a *cut-off* density, 2.5 g/cm^3, high-density plastics, magnesium, and hollow aluminum alloys are recovered in the floating fraction and then processed again utilizing ECS. The "remaining" sink fraction is constituted by brass, zinc, lead, copper, and so on. One of the main limits of S&FS is that it is not capable of performing a good separation of cast from aluminum alloys or differentiating among the different alloy groups. Furthermore, S&FS present other limitations, namely:

- Cost, the process is expensive both at technical (e.g., heavy media costs, complexity of the processing circuit, etc.) and environmental levels (e.g., strict control of the heavy media with reference to possible environmental pollution, water recovery and cleaning, etc.)
- Media recovery, separated scraps are obviously contaminated by heavy media, thus they have to be cleaned and the media recovered (e.g., utilization of specific processing circuits)
- Separation is strongly influenced by particulate solids' morphological and morphometrical attributes

To attempt to totally or partially solve address these issues, in recent years innovative separation/sorting technologies have been successfully proposed and adopted in many recycling plants.

Sorting The main role of sorting strategies is to perform a sort of further refining of non-ferrous metals as they result from the previous separation stages. From this perspective, sorting is mainly addressed at realizing greater discrimination

Recycling Technologies. Table 10 Characteristics of the separation units commonly utilized to perform separation in the non-ferrous metal recycling sector

Separation units	Characteristics
Magnetics Drum Separators (MDS)	MDS is usually constituted by a stationary drum with half of its surfaced lined with NdFeB magnets installed inside a rotating cylinder that is set up as a conveyor belt. Ferromagnetic particles are attracted, removed from the other non magnetic fractions and discharged as the moving belt of the separator turns away from the magnetic field ferromagnetic particles.
Eddy Currents Separators (ECS)	ECS is realized passing the waste products to separate into magnetic field, as a result, eddy current induced in the non-ferrous metals, producing a forward thrust (F) and torque (T) on the particles resulting in their ejection from the stream of non-metallic materials. Separation is realized according to the trajectory followed by the different classes of materials. ECS is commonly applied after the first magnetic separation stage (e.g., MDS). Other eddy current based separation architectures are based on the use of a rotating cylinder or a conveyor belt: magnets are positioned around the rotating axis of the cylinder or fitted inside the head pulley, respectively. As in the previous case separation is realized according to materials trajectory variations.
Air Separators (AS)	AS is usually applied to preliminary recover light fractions contained in the feed or on *non-magnetic-fractions* as they result from previous MDS and ECS. Light fractions (e.g., plastics, rubbers, foams, fibers, etc.) are usually sucked by a nozzle positioned above the conveyor.
Sink and *Float Separators* (S&FS)	S&FS are based on the utilization of a medium constituted by a finely milled solid (e.g., magnetite or ferrosilicon) and water. According to the solid/water ratio the density of the medium can vary. Usually such a value is between the value of the specific gravity of the two classes of materials to separate, so that a sink and a float product is obtained. In this process, the recovered materials are then washed and dried. The fine particles of the heavy media are recovered, by magnetic separation, from the slurry resulting from product washing and re-utilized inside the process. Decreasing the size of the particles to separate, separation efficiency also decreases for the increasing effect of viscosity, in this case cycloning is utilized.

between different families of alloys and different classes inside the families. The technology utilized to fulfill this goal is based on two different approaches: color sorting (**I***mage* **A***nalysis*: IA) and **L***aser*-**I***nduced* **B***reakdown* **S***pectroscopy* (LIBS). For IA, a new process applied to aluminum scrap allows enhancement of IA performance. It consists of a selective etching of the scraps in different solutions that produces, as a result, the coloring of the scrap according to the presence and quantities of specific alloys agent. LIBS allows determination of the chemical composition of each scrap in a reliable and cost-effective way. The main limitation of this approach is linked to the characteristics of the investigated surface (e.g., presence of paints, lubricants, adhesives, or other polluting substances) since the laser pulse laser can only penetrate to a depth of 30 Å or less on the surface of the aluminum. Table 11 lists the devices and related architectures commonly utilized to perform non-ferrous metal scrap sorting.

A new class of sensing devices based on **H***yper***S***pectral* **I***maging* (HSI) has recently opened new interesting scenarios for the *on-line* recognition of the different products resulting from both ferrous and non-ferrous metal waste processing. A more detailed description of this technique is given in the section Future Directions.

Recycling Technologies: Plastics

Re-utilization of waste plastics has increased with "new" plastic polymers; such re-utilization has increased not only quantitatively but also qualitatively (e.g., a larger range of recycled polymers). Plastics can be roughly divided in two types: thermoplastics, which soften when heated and harden again when cooled, and thermosets, which harden by curing and cannot

Recycling Technologies. Table 11 Characteristics of the sorting units commonly utilized to perform sorting in the non-ferrous metal recycling sector, with particular reference to products resulting from S&FS

Detection units utilized for sorting	Characteristics
Image Analyzers (IA)	IA allow to perform sorting on the basis of the detected color. IA belongs to the first class of devices utilized for non-ferrous metal scrap color sorting. Adopting this approach, zinc, copper, brass, and stainless steel are commonly well sorted. Thanks to great technological improvements, both in terms of speed of processing (the same pieces can be checked several times), sensor quality (better discrimination in terms of recognizable colors) and resolution (minimum identifiable scrap piece), recent studies demonstrated that IA allows good separation of magnesium alloys from hollow aluminum products.
Laser-Induced Breakdown Spectroscopy (LIBS)	The detection architecture is based on the analysis of the optical spectrum, or fingerprint, of a small spot on each metal particle that is evaporated using powerful laser pulses. Even if, at least in principle, very powerful, the techniques show some limitations. The most important is that it is particularly sensitive to the status of the scrap surface. Presence can negatively influence measurements because laser pulses can penetrate for just few Å in the surface.

be re-molded. Thermoplastics are by far the most common types of plastic, comprising almost 80% of the plastics used in Europe, and they are also the most easily recyclable. It is easy to understand that when collection-recycling strategies are set up, mixing between thermoplastics and thermosets has to be strictly avoided.

Plastic materials can be considered relatively modern, but some "natural" polymers exist in nature (e.g., amber, tortoiseshell, and horn) that behave very similarly to "modern" manufactured plastics and were used in the past in similar ways. For example, horn, which becomes transparent and pale yellow when heated, was used in the eighteenth century to replace glass.

The first plastic material produced for industrial-scale applications was the Parkesine, later named Xylonite. This material was invented by Alexander Parkes, who exhibited it as the world's first plastic in 1862. It was used for such objects as ornaments, knife handles, and boxes, and for flexible products such as cuffs and collars. Since that time great steps forward have been made and today plastics are widely produced and utilized, creating massive problems with litter and waste disposal.

Plastics are continuously replacing other materials in a number of applications. From greenhouses, mulches, coating, and wiring, to packaging, films, covers, bags, and containers. It is only reasonable expect to find a considerable amount of **P***lastic* **S***olid* **W***aste* (PSW) in the final stream of **M***unicipal* **S***olid* **W***aste* (MSW). In the EU countries, over 250 × 10^6 t of MSW are produced each year, with an annual growth of 3%. In 1990, each individual in the world produced an average of 250 kg of MSW generating in total 1.3 × 10^9 t of MSW [1]. Ten years later, this amount almost doubled at 2.3 × 10^9 t. In the United States, PSW found in MSW has increased from 11% in 2002 [2] to 12.1% in 2007 [3]. Increasing cost and decreasing space in landfills have forced considerations of alternative options for PSW disposal [4]. Years of research, study, and testing have resulted in a number of treatment, recycling, and recovery methods for PSW that can be economically and environmentally viable [5]. The plastic industry has successfully identified workable technologies for recovering, treating, and recycling of waste from discarded products. In 2002, 388.000 t of polyethylene (PE) were used to produce various parts of textiles, of which 378.000 t were made from PE discarded articles [6].

The better solution "to recover" plastic goods should be, when possible, to re-use them as they are. Such a choice can be adopted, for example, for crates or other plastic-manufactured containers, which can be used several times for products and/or materials transportation and handling. Re-using is better than recycling because less energy and resources are required (JCR, 2006). In any case both re-use and/or recycling present several important advantages, related to:

- Reduced use of fossil fuels
- Energy savings
- Reduced emissions of CO_2, SO_2, and NO_x

The main problems in plastic recycling are mainly related to the difficulties of developing and setting up reliable automatic sorting architectures and systems able to perform an efficient selection of the different polymers. These problems can synthetically be divided in two classes, that is: single and multiple type and color plastic separation. In the first case (e.g., bottle, container, etc.) separation is relatively easy, the main problem being related to correct polymer recognition independent of the presence of fillers and other chemical additives. In the second case (e.g., cellular phones, electrical and electronic devices, automotive parts, etc.) separation is more complex, as the objects are constituted by different types of plastics characterized by the presence of different fillers and chemical additives. In the latter case, both separation and recognition strategies have to be sequentially applied.

Waste Plastics Sources and Characteristics Used plastic packaging and other plastic items can be valuable resources in the manufacture of new products and in the generation of energy. It is important that a society aims to make the best affordable use of these valuable plastic resources. This is good for the environment, for the economy, and for the international community. Analyses by the European Community (EU) indicate that besides their ecologic importance, raw materials and energy are also the most important competitiveness factors for EU industries. Therefore, the need to increase recycling, improving at the same time the quality and homogeneity of recycled materials to minimize environmental pollution and usage of resources, is thus an important topic the EU. There is a strong drive to recycle polymers from end-of-life products and avoid their ending up in landfills and waste incinerators because plastics recycling reduces CO_2 emission and saves resources. The worldwide production of plastics was 230 million tons in 2005 [30]. In Europe, 53.5 million tons were produced in total. Out of 22 million tons of post-consumer plastic waste in Europe in 2005, 53% was disposed, 29% was used for energy recovery, and 18% was recycled [30]. According to the last EU Directive 2004/12/EC on packaging and packaging waste, the recycling level of plastics should dramatically increase in the next years. New, more cost-effective separation technology can thus provide an important incentive to increase recycling rates. The recycling of polymers that are present in relatively pure streams such as post-industrial waste and separately collected containers of food and beverage is generally well-developed in Europe. The situation is very different for the large and complex stream of post-consumer waste, including wastes such as ***W**aste **E**lectrical* and ***E**lectronic **E**quipment* (WEEE), household waste, and ***A**utomotive **S**hredder **R**esidue* (ASR). Effective recycling of these wastes is possible, as has been demonstrated at some places in Europe, by large investments in logistics and dismantling (cars, electronic equipment) or hand-sorting (household waste). Such strategies are expensive, however, and they are therefore not widely applied.

There are about 50 different groups of plastics, with hundreds of different varieties [31]. All types of plastic are recyclable. To make sorting and thus recycling easier, the American Society of the Plastics Industry developed a standard marking code to help consumers to identify and sort the main types of plastic. An example of these types and their most common uses are reported in Table 12.

All separation-sorting techniques are based on the identification of one or more physical properties to utilize to discriminate the materials to process in order to establish classes of physical attributes and to set up appropriate technologies to address materials inside these classes. One of the most utilized properties is the density. Unfortunately, such a parameter is not particularly useful when plastics have to be separated because this value is similar for the different polymers to recycle.

Technologies that address these resources need to be extremely powerful, as they must be relatively simple in order to be cost-effective, but also accurate enough to create high-purity products and able to valorize a substantial fraction of the materials, present in the waste, into useful products of consistent quality in order to be economical. On the other hand, the potential market for such technologies is large, and environmental regulations along with oil price increases have

Recycling Technologies. Table 12 Example of most common types of plastics and use

♳	PET	*Polyethylene terephthalate* – Fizzy drink bottles and oven-ready meal trays.
♴	HDPE	*High-density polyethylene* – Bottles for milk and cleaning liquids.
♵	PVC	*Polyvinyl chloride* – Food trays, cling film, bottles for squash, mineral water and shampoo.
♶	LDPE	*Low density polyethylene* – Carrier bags and bin liners.
♷	PP	*Polypropylene* – Margarine tubs, microwaveable meal trays.
♸	PS	*Polystyrene* – Yogurt containers, foam meat or fish trays, hamburger boxes and egg cartons, vending cups, plastic cutlery, protective packaging for electronic goods and toys.
♹	OTHER	*Any other plastics* that do not fall into any of the above categories. An example is melamine, which is often used in plastic plates and cups.

increased the interest in many industries both in waste sorting technologies, for the production of high quality secondary polymers, as well as in developing automatic sensors for quality assessment of waste-derived secondary polymers. This latter aspect is particularly crucial because: "*no matter how efficient the recycling scheme is, sorting is the most important step in recycling loop.*" Fast, accurate, and reliable identification of the primary plastics and the polluting materials in the feed is thus essential to set up suitable mechanical actions on PSW and optimal sorting strategies on the resulting products. The attainment of this goal is thus fundamental for companies that buy recycled plastics, because they obviously want those recycled plastics have the same characteristics as virgin ones. Otherwise, it is not efficient, and sometimes dangerous, to use recycled plastics materials. A simple example is the case of polyethylene theraphalate (PET) and polyvinylchloride (PVC), which are sometimes indistinguishable by sight. These two resins are contaminants to each other. Combinations of PVC and PET resins can result in the release of hydrochloric gases. The PET resin will be ruined even with only a few parts per million of PVC resin.

Waste Plastics Recycling Technologies From what has been previously reported, it is clear that correct plastic recycling is not an easy task. Among all the different wastes materials and products analyzed in this section, plastics are probably the most difficult to separate and recover. Preliminary and efficient plastics sorting, as well as a continuous monitoring of the different waste plastics flow streams, are both key issues to develop optimal PSW products recycling strategies. From this perspective plastics recycling technologies can be divided into four main categories [32]:

- Re-extrusion, that is, the re-introduction inside an extrusion cycle of plastics presenting the same characteristics

- Mechanical, developed to recover different plastic products by a physical processing
- Chemical, addressed to produce feedstock chemicals for the chemical industry
- Energy recovery, that is, complete or partial waste plastics materials oxidation to produce heat, power, and/or gaseous fuels, oils, and/or materials to be disposed of (e.g., ashes)

In the following, particular attention will be addressed to mechanical recycling, one among the four mentioned approaches that maximizes "waste plastics recovery," producing lower environmental impact.

Re-extrusion The main assumption with re-extrusion is that the utilized waste scraps have to be constituted by polymers presenting the same characteristics as the original product. Manufactured products resulting from this process that do not satisfy quality composition constraints are usually addressed to a use where mechanical properties are more important than compositional ones (e.g., crates).

Mechanical Recycling When mechanical recycling strategies are applied, sorting, at the different stages of the processing, represents an important issue. Despite great technological developments, most current plastic sorting continues to be done by hand. *Manual sorting* is a simple process that needs very little technology. It is a labor-intensive, costly, and inefficient method for sorting materials and more specifically plastics. For this reason, as previously outlined, the Society of the Plastics Industry instituted a voluntary labelling system. The system created a set of codes (Table 12) for each of the six most commonly used resin types. Even with this labeling system, it is still difficult to manually distinguish polymer types due to the condition of the plastics as they reach the separation facility. Plastic containers, in fact, may be crushed, cracked, or covered, rendering the resin label practically useless. In any case, systematic and extensive manual sorting of plastic parts, bottles, and the like is counterproductive since accurate, high-speed flake-sorting technology exists to separate one plastic from another; in fact, a number of automated sorting strategies have been investigated, developed, and implemented in recent years. They can be divided in two categories according to the size of plastics objects to sort:

- *Macro-sorting* deals with the separation of bottles or containers, as a whole. Such an approach has the advantage that it does not require any specific preparation of the sample before sorting. Specific polymers' attributes have been detected and according to their characteristics further separated, usually following air-blow-based strategies.
- *Micro-sorting* is applied after the plastic materials have been milled into pieces. This system has the advantages of lower handling costs and larger volume processing. A more sophisticated technology is required: a real mechanical processing sequence (e.g., comminution, classification, separation, etc.) has to be set up and applied.

Macro-sorting Plastic *macro-sorting* is addressed to separate plastics manufactured goods as recovered after their use. Strategies have to be thus addressed to recognize big targets. The main problem to face, following this approach, is to set up a suitable processing line that is able to handle large pieces and consequently large and cumbersome stocks. Different techniques have been investigated in the past years and are currently utilized, some of them are outlined below:

- Near-infrared spectroscopy
- X-rays analysis
- Laser aid identification
- Marker systems

Near-infrared spectroscopy (NIR). This techniques is one of the most utilized to perform an automated sorting of post-consumer plastic containers. NIR has the advantage that direct or close contact between the detector and the sample is not necessary. NIR instruments are also compatible with flexible fiber-optic probes. It is based on the energizing of the unsorted, unidentified plastic with near-infrared waves (600–2,500 nm). When the infrared light reflects off the surface of the plastic, the different resins' characteristic infrared absorption bands can be measured. The detected bands are then compared to known polymer spectral bands response, in the same wavelength range, to determine the plastic type. Such an approach is characterized by many advantages. The most significant one is the detection/identification speed. Because of the great scanning speed allowed by spectroscopic devices, many readings of one sample can be taken in short periods of time; a multiple check of the same object is thus possible, allowing set up of proper and reliable identification strategies. Detection speed also allows an increased volume of plastics sorted in smaller amounts of time. The second advantage is the lack of specimen preparation. Labels or other obstructions like dirt do not significantly interfere with readings thanks to the option of performing multiple checks. Finally this detection architecture presents another advantage: color does not interfere with proper resin identification. Except for black, the readings are independent of the color of the resin. Black containers represent a problem, because their color is a strong absorber in the near-infrared region. As a result, black plastic produces a featureless spectrum that, in many cases, does not allow proper identification [33].

X-rays analysis. This sorting approach is based on the study of the transmitted or reflected wavelengths in the X-ray region. This technology is mainly applied for PVC sorting. Chlorine atoms in PVC give a unique peak in the X-ray spectrum that is readily detectable. X-ray fluorescence (XRF), based on energy level variations of core electrons of atoms, can be used to detect elements in plastics, except for H, C, N, and O, that are usually detectable utilizing infrared spectroscopy. XRF presents many advantages: ease of use, rapid preparation and analysis of the sample, a large range of element detection, etc. It can be thus utilized *on-line*. Furthermore, it allows one to quantitatively determine the presence and characteristics of fillers, pigments, and flame retardants.

Laser aid identification. With this approach the detection architecture identifies plastics by shining a laser beam onto the surface to be identified and then analyzing the material's response. Utilizing an infrared thermographic system, various material properties including absorption coefficient, thermal conductivity, thermal capacity, and surface temperature distribution can be thus determined. The detected properties can then be analyzed to identify plastic type. The resulting system is suitable for quick analysis and identification of various plastics. The approach presents some advantages, that is: (1) different thickness, forms, and surface structures of plastics containers do not play any role in the identification and (2) printing and different additives (softeners) also do not play any role. The limits are related to (1) the presence, in terms of quantity and quality (particularly of carbon), of fillers, (2) difficulty in classification of plastics due to the evaluation of the maximum temperatures directly after laser radiation and, as a consequence of (1) and (2), a lower identification speed, in comparison with spectroscopy and X-ray, wherein checking of plastic containers can be carried out within *only* 1/10 s.

Marker systems. This approach entails marking either the container or the resin itself with something readily detectable. There are no barriers standing in the way of an automated sorting system that would read a hidden marker and identify resin type. Many studies and attempts were carried out in the 1990s, but with low success, mainly due to problems arising both at the production and recycling levels. Every packaging production line would have to install a marking system on their line. Also, each recycler would need to install a machine to scan for the marking on the containers.

Micro-sorting Plastic *micro-sorting* separate post-consumer plastics after a combination of comminution/separation processes specifically addressed to remove contaminants (e.g., non-plastic materials) and to increase bulk density, lowering storage requirements and shipping/transport costs, ease material handling and conveying, and liberating materials. Comminution is thus a fundamental and critical step when complex plastic waste streams have to be processed. Waste feed size reduction, in fact, has to satisfy comminution-liberation requirements and at the same time not produce too much fine fractions, which represent a problem in the further separation stages. According to the recycling plant "input" feedstock, different comminution strategies (e.g., number of comminution stages, utilized equipment, and operative conditions: dry or wet) have to be selected. Table 13 describes some of the comminution units commonly used to perform waste plastics shredding.

After comminution, plastic wastes are reduced in dimension but obviously maintain their original composition, that is plastics and contaminants (e.g., ferrous and non-metals, nonferrous, foams, film, rubber, labels, paint and coatings, metallic foils, glass, rocks,

Recycling Technologies. Table 13 Characteristics of the comminution units mainly utilized to perform waste plastics shredding

Comminution units	Characteristics
Hammer Mills (HM)	Movable hammers mounted on a rotating shaft hit and/or throw plastic against mill chamber or the other waste fed material. As a result comminution is realized. Particles are recycled inside the hammer until they do not reach a size lower than the aperture of a grid installed at the exit of the mill chamber. HM can handle without problems metal contaminants, high energy is required, milled particles are not uniform and the process produces a lot of noise.
Ring Mills (RM)	A RM is usually constituted by a steel rolling blade. This blade chops and grinds the plastic that is placed inside the roller. After it has been ground up to the desired size, it falls through the small holes located beneath the rolling blade.
Shear Shredders (SS)	This machine uses one or more rotating shafts, each with a set of cutting disks or knives mounted closely together on the shaft(s) that sits in a chamber at the bottom of a feed hopper. As the shaft rotates, the cutting devices pull the material down through the small spaces between the cutting disks/knives and the surrounding chamber.
Two or *Four Shaft Shear Shredders* (TSSS or FSSS)	The equipment can be composed by two- or four-shafts shredder with rotary blades (e.g., sharp-corners disks provided with hooks) and spacer combs, that keep the tools clean and make material unloading easy. Once the material goes into the hopper, the shredder catches the material and begins to cut it grossly. Thanks to the high cutting torque and the different conformation of the cutters group it is possible to shred pieces made of different materials. FSSS can handle metal contaminants, relatively low energy is required, particles are well liberated, good size control, throughput rate is lower in comparison with conventional shredding machine without screen, high maintenance cost.
Granulators (Gn)	The main feature of Gn is represented by the rotor conformation provided with short blades with staggered arrangement. During rotation every tool scratches the material and makes the final shredding. Gn are particularly efficient when material characterized by high thickness and resistance has to be cut. They realize a good liberation of the materials, a high throughput, they cannot handle metals (e.g., contaminants) and are characterized by high maintenance costs.
Cryogenic Comminution Units (CCU)	CCU realize a fine grinding by using liquid nitrogen, usually the material is blended with liquid nitrogen to provide sub-zero temperature level up to −150°C, to cool the material in a grinding mill. The cryogenic process produces fairly smooth fracture surfaces. Little or no heat is generated in the process. This results in less degradation of the rubber. Even if the price of liquid nitrogen has come down significantly, recently the process is always characterized by high operative costs. CCU are thus ideal for fine pulverizing of thermo plastic and heat sensitive materials, they also allow to reach an excellent liberation of the materials.

sand, dirt, etc.). Contaminant removal has thus to be carried out adopting different classification/separation strategies strictly linked to the size class distribution of the flow streams and to the contaminants' characteristics, with respect of the polymer/s to recover. For waste plastic recycling, different from what is usually carried out in the recycling sector, some separation stages are carried out in wet conditions, that is, using a fluid, usually water and sometimes heavy-media, to enhance separation devices' efficiency. In some cases, when the feedstock is particularly contaminated (e.g., automotive shredder residue), the recycling process starts with a water or heavy media-based separation stages in order to remove as much contaminants as possible, as metal, rocks, glass, and sand that could damage size reduction equipment could negatively affect the further recycling stages.

To fulfil the previously mentioned classification/ separation goals different techniques are currently used:

- Air classification
- Magnetic and eddy current separation
- Density-based separation processes:
  - Sink-float separation (wet process)
  - Jigging (wet process)
  - Hydrocycloning (wet process)
  - Centrifuge-based separation (wet process)
  - Air table classifiers and gravity table separators (dry process)
- Surface-based separation processes:
  - Electrostatic separation (dry process)
  - Flotation (wet process)
- Selective dissolution

Air classification. Air classification is commonly used to remove, in dry conditions, light contaminants such as dust, small foam particles, paper, glass powders, etc. Usually aspirators or air-cyclones are used [28]. Air classifiers are, in principle, simple devices, their control can vary according to feed characteristics. Equipment has to be correctly set for each stream of material. Separation of material is based on differences in terminal velocities in an airstream and is dependent on particle density as well as morphological and morphometrical attributes. An air-cyclone provides a simple and economical means for most medium to coarse and/or heavier particle collection applications. The centrifugal action and gravitational forces are the operative principles of cycloning. The air flow containing the particles to classify/separate goes through a high-velocity inlet, forcing particles to the collector wall in spiral motion. This, together with gravitational pull, forces the heavier particles downward, while the lighter ones travel upward via the inner vortex and out the air outlet on the top side. Air-based classification represents a fundamental step in any plastics recycling facility, handling complex plastic-rich parts from end-of-life durables (e.g., automotive derived parts, electronic and electrical devices, and appliances, etc.).

Magnetic and eddy current separations. These separation techniques are utilized both at the beginning of the recycling process, as well as after different handling stages. Usually ferrous (e.g., low-grade stainless steel, nickel alloys, etc.) and non-ferrous metals (e.g., aluminum) are removed using magnets [27] and eddy current [17] and/or electrostatic separators [29], respectively. The characteristics of these separation devices are described in the section dealing with metal recycling. The magnetic separators commonly utilized are belt magnets, magnetic pulleys, and drum magnets. The refining/control of the final products is usually carried out by high-intensity permanent magnets.

Density separation processes. Density separation is the most frequently applied technique to recover different plastics from a mixed plastics-pollutants streams. Such an approach can be also be profitably used to separate polymers belonging to the same family but containing different additives. Density-based separation techniques are more reliable than those “only” based on plastics surface characteristics. Bulk plastic properties, in fact, are less sensitive to possible alteration linked to specific environmental conditions (e.g., lighting, oxidation, etc.) or contaminant presence (e.g., oil, dirt, various costing, etc.). When density separation is applied, waste materials to be separated are placed in a medium characterized by a density that is intermediate between two or more densities of the particles constituting the waste. Following this approach, the fluid and/or recovered solid fractions have to be further processed for environmental and cleaning purposes, respectively.

Sink and float process. Sink and float separation systems are very common. They represent a simple and robust approach to separate materials characterized by different densities. The method simply involves depositing the materials in a tank filled with water or other liquid at a specific density. The lighter materials float and the heavier ones sink. For a sink and float system to work efficiently the materials’ densities must differ greatly from one another (e.g., polypropylene, PP: 0.96 g/cm^3, high-density polyethylene, HDPE: 0.94 g/cm^3, medium-density polyethylene, MDPE: 0.926–0.940 g/cm^3, low-density polyethylene, LDPE: 0.915–0.925 g/cm^3, linear low-density polyethylene, LLDP: 0.91–0.94 g/cm^3). Furthermore, even when applied the process is difficult to handle because chemicals have to be added to water, to modify density, or specific heavy liquids [34], as bromoform ($CHBr_3$) (2.87 g/cm^3), TBE: 1,1,2,2-tetrabromoethane $C_2H_2Br_4$ (2.95 g/cm^3) and methylene iodide (3.31 g/cm^3), have to be utilized. Such liquids are highly toxic, require stringent conditions to minimize exposure to workers and create plastics contaminated fraction that have to be further cleaned [34, 35]. Furthermore, the presence of possible contaminants and bubbles on the plastic surface, plastics particle size and shape, and characteristics of fillers and additives also strongly affect separation.

Jigging. Jigging is based on the application of repetitive pulsation actions to a particle bed by a current of water in stratification of plastic waste particles of different specific gravity. A jig operates in a cyclic manner where one cycle consists of four stages, namely, inlet, expansion, exhaust, and compression. In the inlet stage the bed lifts up en-masse. Near the end of the lift stroke the particles at the bottom of the bed start falling resulting in loosening of the bed which, in turn, causes its expansion or dilation. During the third and fourth stages of the jig cycle, the particles resettle through the fluid, and the bed collapses back to its original volume. The pulsation and suction is repeated to bring about stratification with respect to specific gravity across the bed height.

Hydrocyclones. Hydrocyclones are an economical and effective tool for separating mixed plastics and for removing many contaminants from a target plastic. A hydrocyclone transfers fluid pressure energy into rotational fluid motion. This rotational motion causes relative movement of materials suspended in the fluid thus permitting separation of the materials from one another [36]. The mixed fluid enters tangentially at the inlet, which causes the material to rotate within the vessel and ultimately to form a vortex. As this vortex of fluid spirals within the cyclone, heavier materials are forced outward by centrifugal force and down from the barrel section into the cone section. The materials more dense than the fluid flow down the inner wall and exit through the apex and out the underflow port with a portion of the fluid. Lighter materials are swept into the center vortex by inward fluid motion, and are carried vertically up through. Different from classical applications (e.g., mineral processing, food industry, pharmaceutical industry, etc.), when hydrocyclones are utilized in plastics waste recycling, two factors have to be carefully taken into account: (1) the tendency of recycled plastic particles to assume a plate-like shape and (2) the low differences usually existing between different plastics. Hydrocyclones can be considered intermediate density based separation units. For their characteristics, in fact, can be placed between a classical sink-float and a centrifugal process. Hydrocyclones present several advantages: they require very little space, are quite efficient, and outputs can be high; on the other hand, they require a more complex fluid-dynamic circuit (e.g., presence of pumps) and stricter control of feed characteristics (e.g., water solids ratio), etc.

Centrifuges. These equipments are very efficient; they balance optimal separation performances with a reasonably high separation rate. Morphological and morphometrical particle attributes affect in a limited way this separation, because of the applied centrifugal fields characterized by high values. This technique is particularly efficient when fibers and/or film-like particles have to be recovered.

Air table classifiers and gravity table separators. These devices come from mineral processing and metal recycling industries (e.g., automotive derived waste containing plastics). Their application is quite limited.

Surface based separation processes: *Electrostatic separation.* When this separation is applied usually the particle charging method is based on the triboelectric effect. Such an effect is based on a simple principle: when dissimilar materials, for example, particles of two different plastics, are rubbed together they transfer electrical charge and the resulting surface electrical charge differences can be used to separate the two plastics in an electric field; usually charged plastics fall down freely in the area between two electrodes. The particles are drawn to either positive or negative electrode according to the polarity of the charge. According to their trajectory they are thus collected and separated. Many plastics can be separated with this technique: ABS (Acrylonitrile Butadiene Styrene) and HIPS (High Impact Polystyrene) from end-of-life electronic devices, ABS and PMMA (Polymethyl Methacrylate) from automotive waste, PE (Polyethylene) and PP (Polypropylene), PET, (Polyethylene Terephthalate) and nylon, PVC (Polyvinylchloride) and PE from cable scrap, PVC and PC (polycarbonate) from bottles, etc. Electrostatic separation has two main advantages: it can be carried out in dry conditions and the separation architectures and equipments are relatively simple. The main disadvantages are related to the shape of the particles, which influences their surface charge and separation effect. Furthermore when electrostatic separation is applied particles' humidity and moisture have to be strictly controlled.

Flotation. Froth-flotation is another possible method to perform plastics *micro-sorting*. Flotation works similarly to sink and float systems. Froth-flotation is based on the plastic particles surfaces' chemical-physical attributes; for this reason it is particularly suitable for when plastics of similar densities but different surface properties have to be separated. As outlined in [31] a specific plastic can be separated from a complex waste stream by flotation after treating the waste in alkaline solution [32, 33]. Separation of mixed plastic, according to different plastic typologies, can be achieved utilizing appropriate collectors. A large literature exits on this topic [31, 37, 38]. Furthermore, specific wetting agents can be also used to prepare a hydrophobic property [36, 39]. Due to the conditioning some plastics that normally sink (hydrophilic behavior), adhere, according to their composition, to air bubbles generated by a controlled air flow pumped into the system. As a result such particles float to the surface. Materials that are not affected by the bubbles sink to the bottom. Collection systems at the top and bottom of the system can then recover the separated fractions. Other parameters affecting froth flotation are particle size and shape. The main advantages linked to the utilization of froth flotation in plastic recycling are that the technique is well known and settled from a technological point of view and that it is quite flexible in terms of application possibilities. The limitations are primarily related to plastic particle surface status (e.g., dirtiness and/or pollutants) and to the difficulty in defining precise control logics because of plastic waste variability and important factors that have to be taken into account.

Selective dissolution. Selective dissolution is a plastic sorting option that was investigated in depth, as the markers for macro-sorting purposes, in the early 1990s. The process separates mixed or commingled plastics waste into nearly pure reusable polymers without any mechanical pre-sorting techniques. The selective dissolution is based on two different principles: temperature-dependent solubility of different plastics in a single solvent and solvent-dependent solubility of different plastics at a specified temperature. These technologies are not cost-effective for commodity polymers but are sometimes the only solution for the liberation of different coatings associated with PP, such as paint or skin.

Chemical Recycling Chemical recycling, using a depolymerization process, is applied to convert waste plastics, utilizing heat or heat and catalyst, in smaller molecules (e.g., gases, liquids, solid waxes, etc.), that can be used as a feedstock to produce new plastics or other chemical products. The term chemical is thus used, because an alteration is bound to occur in the chemical structure of the polymer. In recent years, a lot of attention has been addressed to this recycling approach (e.g., non-catalytic thermal cracking, catalytic cracking, and steam degradation) in order to produce different fuel fraction from plastic waste products. Several polymers can be profitably processed adopting this approach. Polyethylene terephthalate (PET), certain polyamides (nylon 6 and 6.6), and polyurethanes (PURs) can be efficiently depolymerized. The resulting chemicals can then be used to make new plastics that can be indistinguishable from the initial virgin polymers [40]. Polyethylene (PE) has been targeted as a potential feedstock for fuel (gasoline) producing technologies [41]. The two cited cases are just an example of chemical recycling potentialities. A great deal of literature exists on this topic, because a lot of research efforts and technologies development have been addressed to improve the utilization of this recycling technology. Chemical recycling, in fact, presents, at least in principle, a big advantage with the possibility of treating heterogeneous and contaminated polymers with limited use of pre-treatment [42]. An excellent review and analysis of this technique is reported in Al-Salem et al. [41].

Energy Recovery Energy recovery is based on using waste plastics to produce energy in the form of heat, steam, and electricity. Deriving from crude oil, waste plastics when burned generate a high calorific value. Furthermore, producing water and carbon dioxide upon combustion, plastics behave similarly to other petroleum-based fuels [43]. Such a solution can be considered technically and economically correct when the other recycling strategies (e.g., sorting, mechanical, chemical, etc.) cannot be profitably applied. A typical example is represented by *fluff*, which is the light fine fraction resulting from car dismantling. *Fluff* is constituted of plastics, rubber, synthetic foams, etc. and well fulfils the concept of *waste-to-energy* product. This material, after a "washing" stage, (e.g., polluting materials removal: copper, aluminum, brass, iron, etc.), can be profitably utilized as fuel. Energy production can thus dramatically contribute to increase the full recovery of such a kind of secondary waste with a lower environmental impact (e.g., landfill reduction). Recent studies demonstrated that when plastic waste energy recovery is performed, foams and granules contribute to destroy CFCs and other harmful blowing agents present [44]. However, several environmental problems related to the emissions have to be faced when such an approach is followed, mainly the presence of (1) volatile organic compounds (VOCs), (2) particulate solids, (3) particulate-bound heavy metals, (4) polycyclic aromatic hydrocarbons (PAHs), (5) polychlorinated dibenzofurans (PCDFs), and (6) dioxins. Finally, the presence of flame-retardants (FRs) can influence the combustion process.

All the considerations previously outlined refer to thermoplastics. When thermoset plastics have to be recycled several problems arise. They, in fact, cannot be readily dissolved, melted, re-compounded, and reshaped like thermoplastics. Specific recycling strategies have to be set up. Mechanical recycling is thus primarily addressed to fine grind them for a further re-use as fillers in new thermoset resins or thermoplastic compositions or to recover natural filler or fibers originally utilized in the original thermoset product. In any case, chemical recycling as well as energy recovery processes can be applied to large range of thermoset materials.

Recycling Technologies: Fibers (Textiles and Carpets)

Fibers, both natural and artificial, are commonly utilized in daily life, as well as in technical applications. Also with reference to fiber-based apparels, legislation fixed severe constraints about their disposal at the end of their life-cycle; as a consequence proper recycling systems must be adopted. Recycling can be carried out via two different approaches: (1) to recover energy or (2) to recover fiber materials for their further reuse.

The energetic utilization of end-of-life fibers is not particularly efficient, because the energy generated from burning is less than the energy required for fiber manufacturing. This approach makes sense, from an ecological point of view, since proper combustion results in energy gains without significant air pollution and the consumption of resources. However, the production of synthetic fibers is more expensive compared to non-fibrous plastics. Also the production of natural fibers, like cotton, requires a large use of resources (e.g., water). Hence, product recycling of fibers will increase the sustainability of products and processes. With reference to cotton its caloric value is about 17 MJ kg [45], on the other hand, the energy demand for producing 1 kg of raw cotton is between 38 and 46 MJ kg [45] considering an oil consumption of 1 kg. As a consequence, any recycling process is more convenient than thermal utilization. If the same considerations are applied to man-made fibers, a different fiber-related-energetic-balance can be drawn. The water consumption for the production of synthetic fibers is significantly lower (about 1/10) compared to cotton. For acrylic fibers, for example, the demand ranges between 0.3 and 15 l H_2O per 1 kg of fibers [46]. Energy consumption for polymerization, spinning and finishing is between 369 and 432 MJ kg [45] [46]. Given a caloric value of crude oil between 38 and 46 MJ kg it can be concluded that for 1 kg of fiber about 11 kg of crude oil is necessary. However, during thermal utilization only the caloric value can be used that is about the same or slightly lower compared to crude oil. From this it is obvious that thermal utilization should be replaced by any other process.

Technologies actually available for textile and carpet recycling do not offer satisfying solutions in terms of economic and ecologic demands, this fact is mainly linked to the difficulty of developing a correct separation, first, and a full characterization, after, of fibers, and other polluting materials. Fiber composition as well as their morphological and morphometrical attributes represent important factors to develop optimal re-utilization strategies, this last aspect being particularly relevant in carpet recycling. Recycling technologies have been developed, in terms of logics and complexity, dealing with textiles and carpets, respectively.

Textiles Systematic textile recycling originated in the Yorkshire Dales (Great Britain) about 200 years ago, and the *rag and bone men* of days past were the predecessors of the actual "textile recycling businesses." They collected not only clothing, but also handbags, shoes, bedding, and curtains for re-use. These materials were then often sold abroad, as second hand clothing, but also to provide raw materials to the "wiping" and "flocking" manufacturers and for fibers reclamation to make new garments. Furthermore, it is well known that textiles were and are recycled for papermaking. Textiles recycling can follow different rules according to "recycled textile function":

- The original product function (e.g., clothing reused again as clothes)
- The textile material properties (e.g., absorbency in a wiper, fire retardant non-woven in a mattress spring cover, etc.)

According to their reutilization, recycled textiles can be up-cycled or down-cycled. In the first case they are used for more technically demanding application (higher value); in the second case, they are utilized for less demanding application (lower value). In these two cases, original textiles have to be mechanically processed, adopting specific comminution, classification, and separation strategies in order to recover the constituting fibers from the other materials (contaminants). Products resulting from these approaches are usually: shoddy (e.g., fabric made from the recycling of knitted products), mungo (e.g., fabric made from the recycling of woven products), cotton rag paper made from recycled cellulosic fabrics, etc.

As previously outlined, textile fibers can be classified into natural (e.g., cotton and wool) and synthetic. Recycled fibers demand is strongly influenced by several factors as:

- Fibers composition and characteristics. The presence of fiber blends (e.g., elastic polyurethane) that make recycling more difficult, or the presence of polymers not commonly recycled (e.g., acrylics and polyesters) negatively impact on fiber recycling process
- The possibility to identify new industrial sectors where recycled fiber products can be utilized

Textiles Source and Characteristics Textile wastes can be originated by industry and/or consumers.

Textiles industrial wastes are originated during the processing, production, and/or the manufacturing phase. Such wastes are easy to recycle, the fiber composition and characteristics being known. Contaminants are usually not present.

Textiles consumer wastes are more difficult to recycle. They are usually constituted by fiber mixtures and contain "contaminants" (e.g., non-fibrous materials such as buttons, buckles, or other metal parts).

Waste textiles are usually collected by charitable organization. End-of-life apparel is then sorted according to a possible re-use, that is:

- As re-wearable, cleaning and wiping clothes, short cut for nonwovens as well as for the paper and cardboard industry [45]
- As recovered fibers, in this latter case a mechanical processing has to be applied in order to produce fibers of desired length for their further re-use

Textiles Recycling Technologies According to the considerations previously outlined recycling technologies are applied when "recycled fibers" have to be produced. Waste textiles processing is usually carried out in dry conditions. Such an approach presents two advantages: (1) low energy consumption (e.g., drying is not required) and (2) no water treatment. In this perspective the main actions and the related equipments are reported in the following:

- Human based sorting (e.g., separation of the different apparels typologies)
- Milling and classification, to obtain fibers of the requested morphological and morphometrical attributes
- Separation of the different fibrous and non-fibrous material according to their physical-chemical attributes
- Fiber tailoring and characterization

Manual Preparation and Sorting A human based sensing approach is commonly followed to recognize and preliminary separate the different apparels according to the different types of fibers. After this phase of sorting/grading, clothes are then packed as bales. Each bale is obtained by pressing and identified in terms of average fiber composition and weight. Each bale is thus assumed as the minimal identifiable raw material unit to address to different possible recycling phases, that is:

- Second-hand clothing
- Wiping and polishing cloths for industry
- New products in the reclamation sector (e.g., component for new high-quality paper, upholstery, insulation, even building materials, etc.)
- Filling materials (e.g., car insulation, seat stuffing, etc.)

Milling and Classification Textiles milling is usually carried out utilizing equipment where cutting actions are maximized and the possible comminution effects on hard components (e.g., buttons, zippers, etc.) are minimized to reduce the presence of fine particles contaminants. Classification is usually carried out to recover/remove fine fractions before the further tailoring stage/s. Usually zig-zag classifier and/or pneumatic tables are utilized. As a result of this processing non-ferrous metal and plastics, if present, can also be recovered for further recycling.

Separation Separation is usually addressed to recover the non-textiles materials inside the milled textiles products. Separation is usually carried out adopting magnetic separators (e.g., hump magnet, magnetic pulley, etc.). Metallic fractions (e.g., button, zippers, etc.) are thus recovered.

Fibers Tailoring and Characterization Tailoring is a process specifically addressed to produce individual fibers and to disintegrate all residual textiles and yarns. At this stage of the process the main target is thus to develop a processing sequence able to progressively produce fibers from fabric. A *three-in-one* process, finalized to separate fibers from fabric, is usually applied, that is: picking, pulling, and tearing. Such a goal is usually reached adopting a series of drums with spiked surfaces characterized by an increasing number of finer spikes. Fibers classification is then carried out adopting air classifiers, their characteristics can vary according to classification goals. Tailoring can produce rather long (greater than 2 mm, e.g., new "long fiber" textiles making, nonwovens, etc.) or short (less than 1 mm, e.g., viscosity modification, composite reinforcement, concrete, mortars, adhesives, etc.), fibers according to their re-use. Tailoring process has to be quantitatively and qualitatively assessed, performing a morphometrical and morphological fibers characterization. Fibers morphological and morphometrical attributes (e.g., fiber length, width, and profile structure) thus represent important factors to develop optimal recycled fibers re-utilization strategies, allowing correct identification and new potential applications. Most of the literature describing the recycling of fibers does not provide any details about fiber characterization. Recently, a procedure (MorFi), originally developed for pulp characterization [47], was successfully applied for short fiber characterization [48]. Following this procedure, the length of fibers (*FL*) was measured automatically adopting an imaging-based approach: a suspension flowing through a flat cell observed by a digital CCD video-camera. The analysis of morphological properties of fibers performed by MorFi provided arithmetical average length of fiber (the value most sensitive to the effect of shortening of degraded fibers during their mechanical treatment), expressed by the equation

$$FL = \frac{\sum z_i l_i}{\sum z_i}$$

where z_i is the number of fibers in a given class of length, and l_i is the mean length of fibers in the given class.

Carpets Differently from textiles, carpets represent a more difficult product to recycle, the reason being linked to its compositional characteristics. A carpet, in fact, is usually constituted by a two-layers backing of polypropylene. In between the layers styrene-butadiene latex rubber (SBR) is joined by calcium carbonate ($CaCO_3$) and the fibers are tufted into the rubber (the majority being nylon 6 and nylon 6.6 textured yarns). The SBR adhesive is a thermoset material, which cannot be re-melted or re-shaped. Nylon generally performs the best among all synthetic fibers as carpet face yarn, but it is also the most expensive. This also explains why most of the recycling effort is on nylon recovery. Due to the "complexity" of carpet, at least in comparison with textiles, if fibers have to be recovered, more complex comminution-classification-separation strategies have to be applied. It is possible to recover fibers, mainly polypropylene, from the backing that range from 3 to 25 mm in length. On the other hand, a fraction originating from the pile yarn is obtained. The latter fibers are between 12 to 25 mm in length [49–53]. The composition of this fraction is reported to have 36% PP (fibers from the backing), 18% nylon (fibers from the pile yarn), and 46% (non-fibrous) SBR and $CaCO_3$ [50, 51]. An extensive literature exists on different possible positive re-use of recycled carpet fibers in concrete and soil reinforcement and several other applications [49–53].

Carpet Source and Characteristics Carpet wastes can be originated by industry and/or consumers. Different from textiles, post-consumer-waste carpets represent the larger source. Carpet can be considered a sophisticated product. It is, as previously described, constituted by many materials assembled to assure the durability of final manufactured product, the consequence is that its disassembly and recovery is difficult and requires complex technology and suitable separation control actions to properly recover and certify the different constituents. Carpet-derived products, both fibers and polymeric materials, can be thus originate lower or higher value recycled products, according to the adopted recycling strategies.

Carpets Recycling Technologies As for waste textiles, carpets recycling technologies are carried out in dry conditions. The main processing steps are outlined in the following:

- Preliminary fiber identification and sorting
- Comminution and classification
- Separation
- Solvent extraction of nylon
- Nylon depolymerization
- Melt processing

Fiber Identification and Sorting Carpet recycling strategies have to be set up according to upper surface fiber characteristics. Fiber identification thus represents a key issue to properly address carpet to different downstream recycling tracks [54]. Usually portable infra-red (IR) spectrophotometers are utilized; they usually allow a fast and reliable recognition of nylon 6, nylon 6.6, polypropylene, polyester, and wool fibers. Sorting is usually applied in the collection point; sometimes such a control is centralized in the stocking facilities of the recycling plant. This last solution is usually more efficient.

Comminution Shredding and/or grinding are the commonly applied size reduction actions [45, 55]. These actions are usually performed by a mill with rotary drums equipped with hardened blades. The material after shredding is then sieved by a grid installed at the exit of the mill chamber. The material, which has become sufficiently small, is thus discharged from the mill chamber, the other remains in the mill chamber for re-cutting. Comminution usually produces an increment of the temperature, such a fact can negatively affect milled materials characteristics, for this reason comminution equipments are usually designed to realize a high torque and a low rotational speed. Studies have been carried out to perform carpet cryogenic milling [56]. Following this strategy comminution results are particularly efficient because the *freezing action* of liquid nitrogen, or CO_2, changing the mechanical behavior of the different carpet components allows their better milling and liberation of the different constituents. Costs are usually higher than classical blade based milling and as a consequence the use of this technology is limited. In some cases milling is also realized utilizing water jet [57].

Separation Carpets constituents separation is usually carried out adopting different processing strategies based on a series of combined comminution-classification stages and a further density based separation utilizing both the "simple" gravitational and/or centrifugal field. The second approach is usually adopted to enhance the differences existing between the different materials to separate. Following these strategies filler, nylon and polypropylene can be recovered. A detailed description of this approach is reported in two papers from J. Herlihy [58] and H. P. Kasserra [59]. In some cases the separation of the different carpet components is realized without comminution [60, 61]. Waste carpet is thus subjected to a preliminary clipping of the exposed fiber, a further bombardment of air and steam of the carbonate-filler latex backing and a combination of peeling and picking. The combination of these actions allows recovery of up to 95% of the face fibers.

Solvent Extraction of Nylon Such an approach is utilized when high value nylon has to be recovered from carpet at the end of its life-cycle. Solvents commonly utilized are aliphatic alcohol [62], alkyl phenols [63], and hydrochloric acid [64]. A preliminary comminution of the carpet is always required. The optimal size-class ranges between 1 and 5 cm. Solvents utilization presents both advantages and disadvantages. The advantages are that the use of solvents allows to good recovery of nylon (yield about 90%) characterized by a relatively low degradation. The disadvantages are those related to the use of chemicals and their further recovery and/or re-use, when possible. Furthermore, process temperature and the time required for nylon extraction are other constraints to take into account. Their values change according to the dissolution process adopted. Sometimes **S***uper* **C***ritical* **F***luids* (SCF) are utilized [65, 66]. The process is commonly carried out in batch conditions at high or low pressure and temperatures according to the fluid utilized: CO_2 [65] or formic acid [66], respectively.

Nylon Depolymerization Depolymerization is usually applied to recover nylon, from nylon carpets, because nylon resin has a remarkable higher value than the other polymers commonly utilized in carpet manufacturing [67]. A typical depolymerization process [68] is based on a preliminary carpet sorting and a further mechanical shredding, the recovered nylon six face fibers are sent to a depolymerization reactor and treated with superheated steam in the presence of a catalyst to produce a distillate containing caprolactam, that is the single monomer that after polymerization originates nylon 6. The crude caprolactam is then distilled and re-polymerized to produce again nylon 6 [69]. Another process to obtain caprolactam is based on the utilization of a two-stage pyrolysis process. The ground nylon carpet, without separation, is dissolved with high-pressure steam and then continuously hydrolyzed with super-heated steam to form caprolactam [70, 67].

Melt Processing Melting and compounding are two other processes currently carried out for carpet recycling. Both require a preliminary size-reduction process. By melting, a thermoplastic polymer is converted by extrusion in resin pellets [45, 55], if more polymers are blended together and then extruded a compounding process is applied. For its characteristics, the products resulting from comminution have to be “compacted” before the extrusion process; they, in fact, are quite bulky. Specific equipment (crammer-compactor feeder) is thus utilized. Extruders can vary, ranging from single screw, twin-screw co-rotating, or twin-screw counter rotating architectures according to feed and required products characteristics. The melted-extruded products are then cooled and cut to obtain pellets. Usually ring, strand or underwater pelletizers are utilized. In Table 14 are synthetically reported some of the characteristics of pelletizing devices commonly utilized for waste-carpets-based melted-extruded products chopping.

Melted-extruded products present different characteristics and market values according to the characteristics of the original feed-stocks [71]. When carpets are constituted by polymers as plastic and polypropylene [72], the resulting pellets (e.g., compounds) are of low quality due to the fact that they are thermodynamically unstable when melt-mixed [73, 74]. They must be thus stabilized to prevent coalescence during melt processing [75]. This process of stabilizing polymer blends is commonly called *compatibilization*. It usually consists of the addition of a premade block copolymer composed of blocks that are each miscible with one of the homopolymers [76].

A melting plant is relatively simple to utilize and maintain. Its main limits are primarily related to its operating principle, that is, low flexibility in terms of possible modification of final products characteristics. Furthermore, the presence of water represents a further environmental processing problem (e.g., H_2O filtering, temperature control and re-circulation to face when melting is applied).

Pellets resulting from the previously described melt-based process are commonly utilized in the

Recycling Technologies. Table 14 Characteristics of the pelletizing equipment utilized for waste-carpets-based melted-extruded products chopping

Equipment	Characteristics
Water Ring Pelletizer (WRP)	A WRP is commonly utilized for liquid polymers in which pellets fall from cutter knives into an annulus on the surface of a body of cooling liquid. The velocity and trajectory of the pellets are controlled by the projection of a spray of cooling liquid radially across the dies from which the pellets are severed. Band heaters in proximity to the dies maintain the polymer in a liquid state prior to extrusion. WRP is particularly suitable to be applied for polymers characterized by a low melt flow index (e.g., polyethylene).
Strand Pelletizer (SP)	In a SP polymer strands discharging from the die head are sent to cooling water-based-device. The water "wetting" the polymer strands is eliminated utilizing an "air knife." The dry, solidified polymer strands are then delivered towards a strand pelletizer where cutting is applied. One of the main disadvantages of SP is that a large floor space and particular care has to be addressed to control possible strands breakage. Polymers as nylon, polyester terephthalate, and polypropylene are commonly processed by SP.
UnderWater Pelletizer (UWP)	UWP is usually constituted by an extruder that conveys the polymer melt to the die plate through the start-up valve. The melt stream is then divided into a ring of strands that flow through the annular die into a cutting chamber flooded with process water. A rotating cutter head in the water stream cuts the polymer strands into pellets, which are immediately conveyed out of the cutting chamber. The pellets are cooled and transported in a slurry to the centrifugal dryer. Pellets are then separated by water through rotating paddles. Polymers as nylon, polyester terephthalate, and polypropylene are commonly processed by UWP.

molding process, as they result from recycling or are blended with virgin polymers. Other common applications are in glass fiber-reinforced composites, where they are utilized as matrices [77].

Future Directions: Innovative Control/Sorting Devices/Logics Integration in Recycling Plants

The different recycling technologies described and analyzed for the different waste materials (e.g., paper, glass, metals, plastics and textiles) clearly demonstrate that improvements in comminution-classification strategies, and further separation technologies (e.g., magnetic, electrostatic, sink-float separation, flotation, etc.) can be mainly carried out by introducing innovative control devices and architectures, as equipment technology and characteristics have reached a very high level of quality and reliability. Such classes of innovative devices can be also utilized as detection systems to realize innovative sorting architectures. Single and/or combined control/sorting actions can be thus developed, taking into account different aspects: (1) waste streams physical-chemical characteristics, (2) market requirements for concentrates of specific quality, and (3) related innovative control/sorting device/logics operating at different scale, that is, at single equipment and/or plant scale.

Waste Feed Streams Characteristics

Processing actions applied to different particular solids waste streams, constituted by different materials (e.g., paper, glass, metals, plastics and textiles), usually perform a change of some physical attributes of the wastes, these changes depend on the intrinsic characteristics of the constituting materials and the actions applied. For example, comminution produces a reduction of the size class distribution of waste, originating smaller particles of different morphological and morphometrical characteristics and producing the liberation of the different materials originally locked (e.g., mixed particles). On the other hand, separation actions allow grouping together of particles according to a specific property (e.g., density, conductivity, magnetic, color, texture, etc.), originating a concentrate.

What are the strategies to apply to make improvements in terms of a correct utilization of waste streams characteristics versus *adopted recycling technologies finalized to a higher recovery of concentrates?*

The answer is in principle very simple, being related to the correct application of recycling technologies taking fully into account waste streams' physical-chemical characteristics. For example, the definition of proper comminution strategies addressed to obtain adequate size class distribution, minimizing the presence of fines and ultrafine particles, is obviously related to a correct knowledge of wastes. With reference to plastics, milling actions based on cutting represent the best solution; on the other hand, impact actions better realize comminution for glasses. Lack of knowledge of the composition of the waste materials to process, in terms of constituents and their time variation, can strongly impact the quality and quantity of the final recovered products. Waste materials present a high degree of compositional variability. Batch sampling, as usually performed on feeds and processed flow streams, does not allow performance of a full and continuous monitoring of the materials flow. Low-cost, reliable, and robust waste streams' physical-chemical characteristics detection devices, realizing continuous monitoring, could represent the solution to performing a *full-time-independent-evaluation* of waste materials streams handled in the plant.

Waste Derived Concentrates Recovery and Quality Assessment

Concentrate quality affects its value and, as consequence, the economic revenue of the recycling process. The possibility to realize a continuous full monitoring of concentrate characteristics is important: (1) to apply correct waste processing strategies to quantitatively and qualitatively maximize recovery, (2) to build a production data base embedding produced product characteristics, and (3) to perform a full products certification.

Recovery maximization is one of the key issues when recycling technologies are applied. Waste products are usually constituted by different materials of different characteristics. An optimal target could be represented to set up recycling actions addressed to separate and recover all the different waste constituents. In this case "the zero-waste" target should be fully reached. If successful, this strategy could allow re-integrating all the wastes in new production cycles. Such a goal is almost impossible to reach, but the strong scientific development, the related technological innovation and the larger attention of new generations towards environmental problems has and will continue to contribute to introduce new technologies for a more efficient recovery. To maintain a trace of when and what is produced in terms of WDPs is another key issue. The achievement of this goal allows establishment of a time correlation between waste feed and resulting concentrate. Correlations are useful not only in technical terms, for a better understanding of plant behavior with respect to waste feed variations, but also because they provide useful information about consumption and related waste production, contributing this way to better waste collection and handling strategies prior to the application of recycling technologies. Finally, product certification in the recycling sector sometimes represents a negative point. The proposed approach could strongly contribute to solving this problem, allowing at the same time a big step forward in product quality detection performed continuously and not on a batch basis.

Control Actions and Logics

As outlined in the previous paragraphs, one of the key points related to a systematic introduction of control logics, inside processing layouts, is to define simple, reliable, robust, efficient, and low-cost architectures able to perform a full characterization of the different flow streams in terms of waste feed and/or resulting product composition, that is, grade of the recovered materials and presence and characteristics of pollutants. The two aspects are intimately linked to the definition of suitable *process-control-strategies* and to the subsequent certification of the recovered materials.

How can control actions and logics be fully introduced and widely utilized in waste recycling?

Each material is characterized by specific attributes; these attributes are usually detected, with sensing devices able to collect one or more piece of information related to the characteristics of handled materials. Information is then processed following logics oriented to maximize the positive effects of the single and/or or group of actions. Control actions can be commonly categorized in four groups:

- *Feedback control*: a control system that monitors its effect on the resulting product. On the basis of the collected information, tuning actions are applied on equipment, recycling plant sections, and operative variables in order to modify the output (product) accordingly.

- *Feedforward control*: a control in which changes are detected at the process input (waste feed) and an anticipated correction signal is applied before process output (concentrate) is affected.
- *Cascade control*: an automatic control system in which various control units are linked in sequence, each control unit regulating the operation of the next control unit in line.
- *Ratio control*: a control procedure in which a predetermined ratio between two or more variables is maintained.

Feedback and feedforward controls are most commonly utilized in recycling.

Waste recycling flow streams are usually constituted by complex particulate solids systems. A particle is thus the minimum portion of material that can be processed. Each particle is characterized by different attributes: size, shape, composition, texture, etc. To define a control logic means identifying one or more rules to handle one or more of the previous mentioned attributes as they have been collected inside a particle flow stream, with the specific aim of verifying whether processing has produced an output satisfying the expected requirements of these attributes, both qualitatively and quantitatively, in the concentrate. In plastic recycling, for example, if a recovery of PP and/or PE has to be carried out, the presence of other plastics materials as PET or PVC, or other pollutants, as metals, glass, etc., must be avoided. In this case, the attribute composition plays a pre-eminent role in control logic definition. Actions to perform are related to: (1) the characteristics of equipment/s generating the PP or PE concentrate and (2) the parameters allowing their operation. Relationships existing between equipment/s operative conditions (processing parameters set up) and quality of the output, for a specific feed, have to be clearly investigated and formalized, constituting the basis for control logic implementation. A quantitative evaluation of the composition of PP and PE concentrate is thus fundamental to set up the logic in a quantitative way. In this case, as when recycling technologies have to be applied to other waste products, the quantitative collection of the attributes qualifying, or certifying, an intermediate and/or a concentrate product is not easy in terms of economically acceptable devices. This problem is common in waste recycling, where the value per unit of weight of product is usually low. To be economical, a process requires the processing of large quantities of waste and the production of a corresponding high amount of waste-derived concentrate. Often this condition does not match with the adoption of sensing devices that, to fulfill the previous requirements, are usually very expensive. A different approach has thus to be followed. If composition can be correlated with other parameter/s, easily detectable, a transposition logic can be applied: that is, the correlated property can be assumed as the control parameter and new control logics applied. Such an approach is particularly meaningful when *on-line* control logics have to be defined. The possibility to apply transposition logics sometimes produces a strong simplification in control. An example of what is described is reported in the following where different examples of sensing architectures based on **H***yper***S***pectral* **I***maging* (HSI) techniques are reported with reference to some of the materials taken into account in this section.

Plant Scale Control

The application of control actions and logics, on a recycling plant scale, can be considered a common practice. Originally performed, and in some cases also today applied, following a human senses-based approach, with the technological development "humans" have been and are being replaced by sensing devices. The first step to implement a plant scale control consists of the identification of some key points inside the plant where product characteristic detection can give useful information about the process. According to the values of the detected parameters, and utilizing pre-assigned rules, the equipment operative variables are modified accordingly. These actions can usually be performed with a feedback or a feedforward approach, more rarely adopting a cascade control. Approaches vary according to waste materials characteristics, control objectives, and implementation modalities: (1) introduction of control logics inside an existing plant or (2) in new ones.

In this latter case, a large flexibility in control architecture design can be performed, not existing predefined plant architectural constraints.

Single Equipment Control

Following this approach, control actions and logics play at single equipment scale. Process and product parameters detection is carried out before (feedforward) and/or after (feedback) the single processing unit and control logics act accordingly. Such an approach is very promising, especially for future improvements in the recycling sector. The "intelligent processing machine," is equipment able to "understand" how it works, according to feed flow stream and resulting product characteristics, and able to modify its "behavior" accordingly, adopting preassigned and/or time-dependent learning rules. This is particularly relevant in recycling, where very often handled materials do not show constant compositional characteristics, affecting equipment performance.

Innovative Sensing Technologies in the Waste Sector

In recent years a lot of innovative sensing technologies and related control logics have been proposed in recycling. A sensing station is usually constituted by a conveying device (e.g., conveyor belt, vibrating channel, etc.) for the separation and steadying of the material, and a detection unit, positioned underneath or above the conveying device or at the material discharge area. Collected information can be then utilized to quantitatively assess material characteristics (certification) or to modify operative conditions of the equipment handling the materials before or after sensing (control) or to develop separation actions by actuators, e.g., valve bank blowing out materials constituents according to specific component characteristics (sorting). In the following, some sensing devices and their operative principles, together with possible application fields, are described.

E*lectronic* I*maging* (EI) Vis*ible* (VIS) Wavelength Based EI-based sensing devices and algorithms [78] are the most widely used in recycling. They belong to the first class of control device utilized in this sector. EI application in recycling technologies comes from architectures developed in mineral processing and food control sectors. Such a technology moves from black & white (B&W) to color sensing systems. Actually color is widely applied in many recycling sectors: glass, WEEE, metals scraps, fluff, wood, etc. Color-based sensing is based on the collection and analysis of materials surface characteristics. Such an approach can represent sometimes a limitation because the collected information "only" relates to the surface (e.g., varnished objects in principle cannot be detected in a material-related way). In many cases, for the relatively low cost of this approach, EI will solve many control/sorting/quality assessment problems when the materials to check have passed shredding stages beforehand that remove the existing surface coatings or that break up the material in a way that with utmost probability uncoated fracture faces can be observed [79]. EI-based detection architectures are commonly constituted by an energizing source (lighting system), a CCD, matrix, or single array camera. Single array-based devices are particularly suitable to be utilized in waste material quality control because the detection principle (scan line) and the target of investigation (waste particles) moves towards each other with a constant speed. Waste particles can thus be fully investigated adopting different time-scale-related sampling strategies according to control/quality actions to apply.

E*lectronic* I*maging* (EI) n*ear* I*nfrared* (NIR) Based The principle that is at the basis of NIR technology is the measurement of object reflectivity within a wavelength ranging between 1.100 and 2.100 nm [80]. In this wavelength range, materials such as plastics, paper, and textiles are characterized by specific spectral firms allowing their recognition. This range of wavelengths is not visible to the human eye. This is also the reason that optical sorting systems must be used. The new generation of NIR-based detection devices can operate a good plastics distinction (e.g., PP, PS, PET, EPS, PC, or PVC) [81], as well as allow identification of materials such as paper, card, cardboard. or wood and natural fibers. A recognition limit of such a procedure is in the identification of materials that cannot be identified due to a lack of individual stone, porcelain, or dark materials, that, for the low level of reflectance, in the investigated wavelength range, do not allow recognition.

X-*Ray* F*luorescence* S*pectroscopy* (XRF) XRF is based on X-ray emission by a radioactive source or X-ray tube inside an XRF unit. X-rays emitted are absorbed by the atoms in the waste sample generating fluorescence as the atoms relax and release energy. The emitted wavelengths, as well as the energy released, are a function of the elements constituting the waste sample. Emission intensity is correlated to the content of a specific element within the sample. Such a technique was and is widely used in many recycling sectors as: metal/alloys, glass, plastics, wood (detection of arsenic, chromium, and copper in treated wood [82] and treated wood waste [83–85]), WEEE, waste-derived-fuels, etc.

D*ual* **E***nergy* **X-R***ay* **T***ransmission* **(DE-XRT)** DE-XRT is similar to that applied for luggage inspection and medical applications. Waste products are transported by a conveying belt. Waste material is energized, from the bottom, by X-rays. Transmitted radiation is collected by X-ray line detectors [86]. In order to separate the effects of density of the X-rayed object and the material thickness, the radiation intensity is generally measured in two different energy ranges. Following this approach thickness influence is eliminated and the radiation that passes through the material allows performance of materials recognition according to its density. The sensing approach allows classification of the material on volume basis. Different from EI or NIR techniques, surface characteristics (e.g., dust, water, small quantities of pollutants) do not practically influence recognition.

L*aser-***I***nduced* **B***reakdown* **S***pectroscopy* **(LIBS)** The system analyzes the optical spectrum, or fingerprint, of a small spot on each metal particle that is evaporated using powerful laser pulses. LIBS has been widely used as a diagnostic technique for the analysis of both surfaces and gaseous streams for metals [87, 88]. LIBS uses a high-power laser that is directed towards the sample through a series of mirrors and lenses that create a small micro-plasma on the surface of the targeted material. Due to the extremely high temperatures produced within the plasma (greater than 20,000°C) the atoms within the plasma emit light (or energy) characterized by different wavelengths. Certain wavelengths of energy are unique to different elements. The system then analyzes the optical spectrum, or fingerprint, of the micro-plasma collected and re-directed toward a fiber optic cable, which then feeds the signal to a spectrometer. The intensity of the emission is directly proportional to the amount of that element present in the sample. The costs and complexity of the system, as well as limitations in efficiency, are the main reason why LIBS is only applied in a number of very specialized operations, commonly for surface contamination. Applications have been developed with reference to wood waste contaminated with chromated copper arsenate sorting [89] and scrap metals [90].

H*yperspectral* **I***maging* **(HSI)** HSI, known also as chemical or spectroscopic imaging, is an emerging technique that combines the imaging properties of a digital camera with the spectroscopic properties of a spectrometer able to detect the spectral attributes of each pixel in an image. Thus, a hyperspectral image, is a three-dimensional dataset with two spatial dimensions and one spectral dimension.

HSI was originally developed for remote sensing applications [91], but has found a large utilization in such diverse fields as astronomy [92, 93, agriculture [94–96], pharmaceuticals [97–99], medicine [100, 101], and in recent years in the recycling sector [24, 102, 103], where important projects were also sustained by the European Union [104, 105].

Among the previous mentioned sensing techniques HSI can be considered one of the most interesting and subject to a wider and wider utilization inside the recycling sector. HSI, in fact, presents some advantages related to its intrinsic characteristics:

- Continuous monitoring, with HSI-based devices as scan line cameras,
- Utilization of different time-scale-related sampling strategies, according to the control/quality actions to develop,
- Implementation of fast and reliable recognition logics, strongly linked to HSI detectors characteristics (e.g., possibility handle spectra as images),
- Null environmental impact of the device,
- Relatively low costs of the devices.

Furthermore, HSI devices, and related operative architectures can be easily integrated inside existing recycling plants, or implemented in new ones, with an optimal cost-benefit ratio. HSI, for its intrinsic properties, can thus be profitably utilized, both as a smart detection engine for sorting and as flow stream quality control, that is, certification of recovered materials and/or products. HSI is fast, accurate, affordable, and it can strongly contribute to lowering the economic threshold above which recycling is cost efficient. For its characteristics it can be meaningfully and reasonably developed, and applied, with reference to many solid waste-handling-sectors ranging from inorganic to organic waste sources.

Different cases studies describing the potentialities of HSI integration inside recycling technologies are reported in the following section.

H*yperspectral* I*maging* (HSI) Based Applications

An HSI system is typically constituted by optics, a spectrograph, a camera, an acquisition system, a translation stage, an energizing source (lighting device), and a control unit (PC). The camera, spectrograph, and illumination conditions determine the spectral range of the detection architecture. The sample/target is usually diffusely illuminated by a tungsten-halogen or LED source. A line of light reflected from the sample enters the objective lens and is separated into its component wavelengths by diffraction optics contained in the spectrograph. A two-dimensional image (spatial versus wavelength dimension) is then formed on the camera and saved on the computer. The sample is moved past the objective lens on a motorized stage and the process is repeated. Two-dimensional line images acquired at adjacent points on the object are stacked to form a three-dimensional hypercube that may be stored on a PC for further analysis.

The applications described in the following are based on sensing devices working in two different wavelength spectral ranges, from 400 to 1,000 nm (VIS-NIR range) and from 1,000 to 1,700 nm (NIR range). The first consists of a CCD camera, a line scan spectrograph (ImSpector™ V10E, SpecIm™, Finland), a lighting architecture, the spectrograph ImSpector™ V10E operates in the spectral range of 400–1,000 nm with a spectral resolution of 2.8 nm. The details of the acquisition architecture are reported in Table 15. The second is a SpecIm NIR spectral camera consisting of an ImSpector N17E imaging spectrograph for the wavelength region 1,000–17,000 nm and a temperature-stabilized InGaAs camera and a lighting architecture (Table 16).

Recycling Technologies. Table 15 Technical characteristics of the ImSpector™ V10E

Sensor	– 2/3″ CCD Array 780 × 580
	– Firewire digital output
	– Pixel resolution: 12 bit
Spectral range	400–1,000 nm
Spectral resolution	2.8 nm
Smile	< 1.5 µm
Keoneyst	< 1 µm
Entrance slit	30 µm × 14.2 µm
Image size	6.5 mm × 14.2 mm
Numerical aperture	F/2.4
Illuminant	– Anodized aluminum cylinder – Barium sulfate internal coating – d/O illumination and viewing conditions – Adjustable height and distance – 150 W cooled halogen lamp
	– Stabilized power source

Both equipments are installed to perform the inspection of the waste materials on a laboratory scale conveyor belt (Fig. 2). The two devices are fully controlled by a PC unit equipped with the Spectral Scanner™ v.2.3 [106] acquisition/pre-processing software.

Sample Set Selection The waste or derived products investigated belong to some of the classes of materials analyzed in this Section and characterized by different specific attributes, different sorting-selection problems, and quality requirements. From this perspective, waste glass fragments (*cullet*), light fraction derived from car shredding residues (*fluff*), and complex plastics-based waste streams have been selected. Results show that the HSI approach allows development and set up of strategies able to reduce analytical costs, improving the speed of the waste streams characteristics detection/analysis, and/or simplifying the

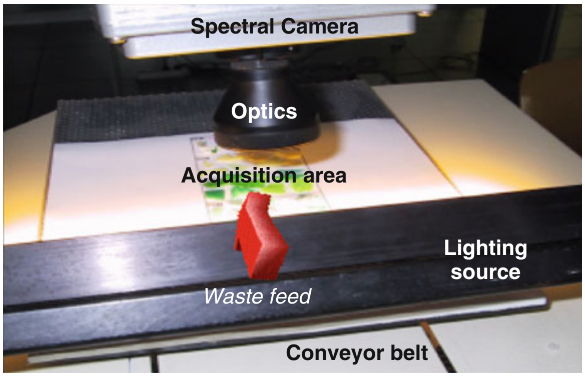
Recycling Technologies. Table 16 Technical characteristics of the ImSpector™ N17E

Sensor	– TE-cooled INGaAs photodiode array 640 × 512 – 14 bit, USB$_2$, LVDS, CameraLink
Spectral range	900–1,000 nm ± 10 nm
Spectral resolution	2.6 nm
Spatial resolution	Rms spot radius < 15 μm
Aberrations	Insignificant astigmatism, smile or keystone
Effective slit length	12.8 mm
Numerical aperture	F/2.0
Stray light	< 0.5% (halogen lamp, 1,400 nm notch filter)

procedures in terms of possible *on-line* implementation of fast and robust classification procedures oriented to develop innovative control strategies "human judgments and error free" as well as innovative certification criteria.

Spectra Acquisition and Detection Logics Implementation Spectra related to the different investigated can be acquired adopting the acquisition architecture described in Fig. 2. Such a strategy is adopted because it mimics, at laboratory scale, the real behavior of the control architecture at an industrial scale, that is, the progressive and continuous horizontal translation of the sample and the "synchronized" acquisition of the spectra at a pre-established step, allowing a tuning of the detection/inspection frequency of the waste materials according to their characteristics. Analyses can thus be performed to verify the fulfillments of different goals:

First goal: the possibility to identify specific spectral attributes for each of the constituents of the different waste streams according to their intrinsic chemical-physical characteristics. Starting from this information, analysis can be carried out performing a characterization of the "shape" of the entire detected spectra and/or identifying, at specific wavelengths, peaks or valleys characterizing the detected spectral firm.



Recycling Technologies. Figure 2
Particulars of the architecture set-up utilized to perform a progressive and continuous waste samples spectra acquisition based on the ImSpector™ series devices

Second goal: the definition of fast, reliable, and robust recognition-classification procedures, based on different logics as: (1) spectral firms correlation, (2) single band intensity comparison at specific wavelengths, and (3) specific wavelengths intensity ratio analysis, in order to perform the discrimination of the different constituents inside a specific waste stream and to allow to reach a certification of the different products in terms of their composition.

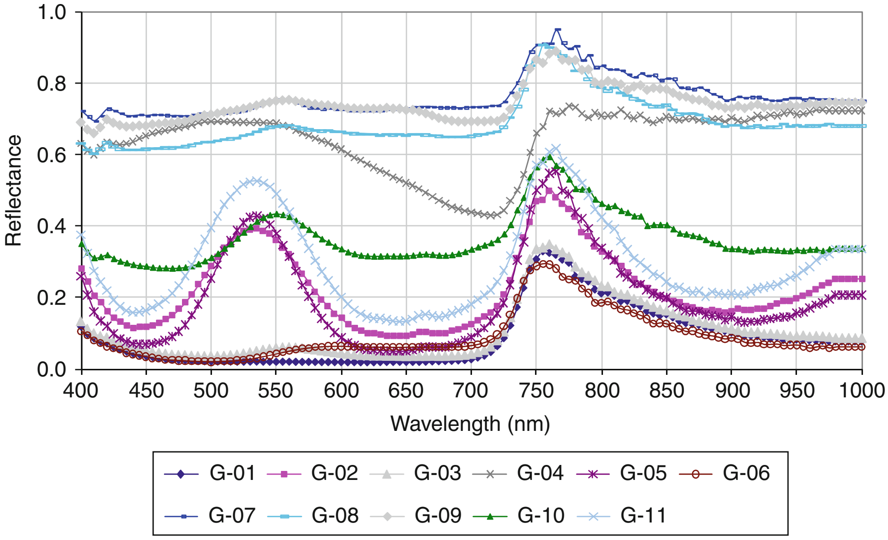
Third goal: the possibility to perform a correlation among detected spectra, sample textural attributes, presence, characteristics, and localization of "pollutants": this latter aspect being of great interest to develop innovative sorting strategies. To validate the efficiency of the HSI-based technique to perform a topological assessment of the different materials an approach based on Principal Component Analysis (PCA) can be also adopted.

Case Studies Tests reported are referred to different solid waste materials characterized, as previously outlined, by a different nature and physical-chemical attributes and "affected" by different processing-separation and/or control problems. For each investigated waste product, a synthetic overview of the target to reach by the HSI approach (*Issues*) and current status of the related utilized characterization approaches (*State of the art*) is reported.

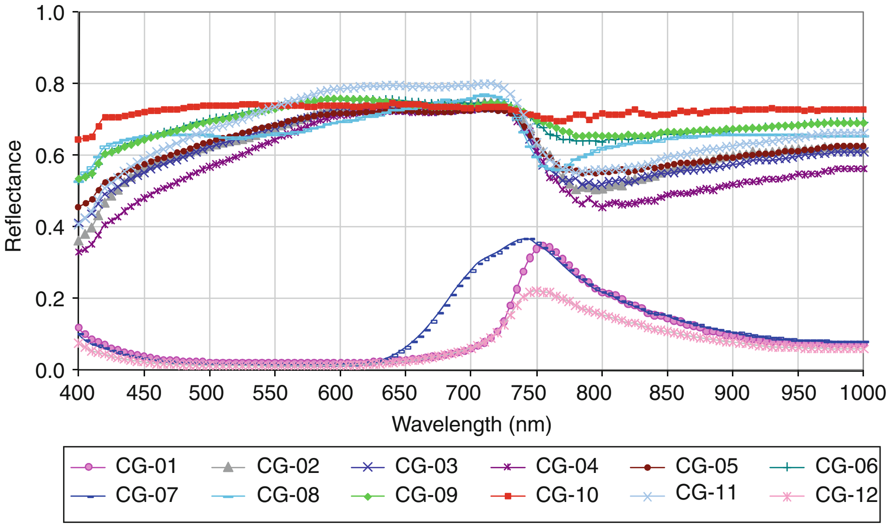
Case Study No. 1: Ceramic Glass Identification Inside Waste Glass Products (cullets). *Issues*: The amount of ceramic glass in post consumer glass waste stream is strongly increased in recent years, mainly for the introduction on the market of many ceramic glass manufactured goods [15]. These products constitute a new generation of consumer household goods used for their thermal-shock resistant properties. It can be argued that, considering the typology of ceramic glass products, contamination involves both the main production lines of glass recycling plants, e.g., the flat glass cullet and the container glass cullet. Due to its physical properties, similar to those of glass, ceramic glasses are almost impossible to detect adopting the automated optical technologies, commonly utilized for cullet color sorting [107]. Therefore, identifying and removing ceramic glass from the glass waste stream has long been a challenge for recyclers of glass.

State of the art: The two actions usually carried out to realize transparent ceramic contaminant removal are source reduction and manual sorting. Therefore, there is the need for the development and the implementation of a system able to realize a real-time identification of ceramic glass in the cullet stream. Sorting techniques based on X-ray [79] and FT-IR spectroscopy [24] have been proposed as methods for ceramic glass detection, but both are expensive and difficult to implement for safety and efficiency reasons, respectively.

The HSI approach clearly has potential in developing innovative sorting strategies. In the visible range (400–700) nm the technique presents a high discrimination power for classical cullet separation by color, allowing the possibility to decrease the minimum size of sorted glass particles, on the other hand, in the visible range the recognition of glass from ceramic glass is almost impossible. Moving in the wavelength range between (700–1,000) nm and (1,000–1,700) nm it is possible the architecture allows discrimination between glass (Fig. 3) and ceramic glass fragments (Fig. 4). Such a discrimination is relatively easy for clear samples; darker samples are more difficult to discriminate. Color, in fact, seems to affect the reflectance levels. Amber glasses show reflectance values near zero, green glasses present intermediate reflectance values, whereas clear glasses display higher reflectance values. A two-steps sorting, a first based on HSI acting between (700–1,700) nm and the second in the (700–1,000) nm range, could thus allow good preliminary *cullet* sorting targeted to ceramic glass contaminant identification/removal and a further *cullet* color-based separation. Application of this technique, in the waste glass sector, can thus allow development of innovative



Recycling Technologies. Figure 3
Spectral plots in the VIS–NIR field (400–1,000 nm) of glass samples



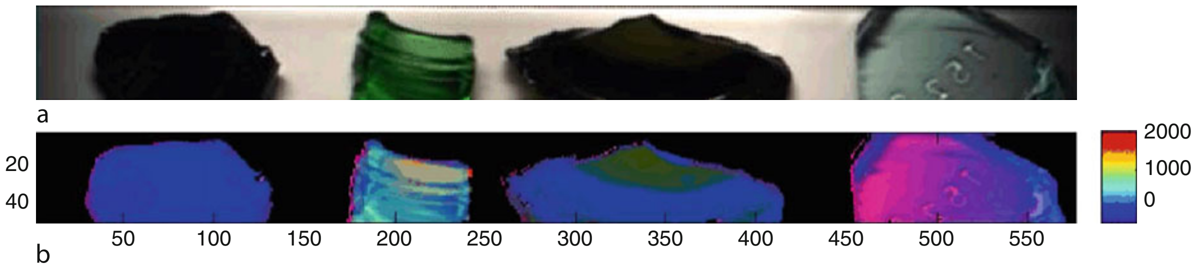
Recycling Technologies. Figure 4
Spectral plots in the VIS–NIR field (400–1,000 nm) of ceramic glass samples

sorting logics: (1) specific attributes can be associated to particulate solid materials, (2) classification procedures based on wavelengths intensity ratio can be defined to perform sorting, and (3) the elements to sort can be topologically assessed adopting a Principal Component Analysis (PCA) [108]. An example related to this last point is outlined in Fig. 5. The HSI devices for their hardware architecture can be easily installed both to integrate and/or to substitute classical sorting imaging-based devices.

Case Study No. 2: Light Fraction Resulting from Car Dismantling (Fluff) Characterization. *Issues*: *Fluff* represents about 25% of the weight of a car and is usually constituted by materials characterized by intrinsic low specific gravity (i.e., plastics, rubber, synthetic foams, etc.). When processed to perform its recovery, *fluff* is polluted by materials presenting higher specific gravity (i.e., copper, aluminum, brass, iron, etc.), constituting parts of the electrical devices of the vehicle that, for their shape, size (i.e., wires, metal straps, slip rings, wipers, etc.) and utilization remain concentrated in the lighter products. Such “polluting materials,” for their intrinsic characteristics, are not well removed by classical separation techniques.

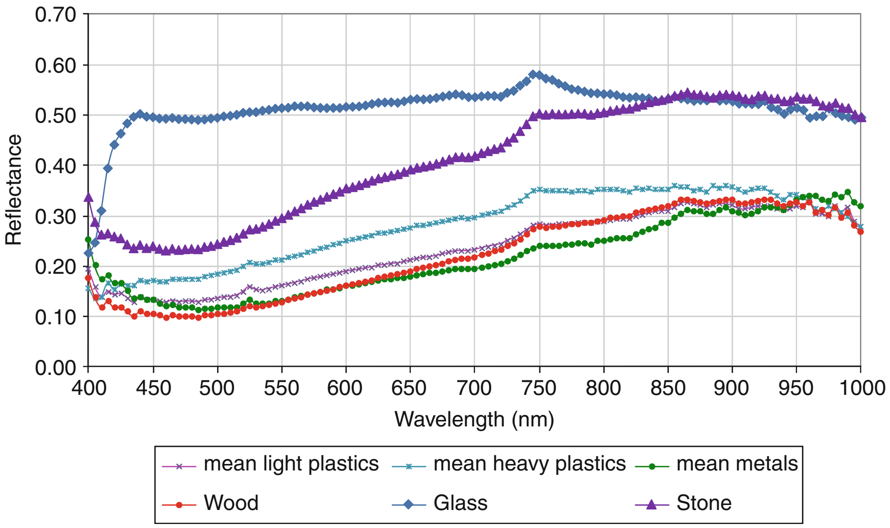
State of the art: *Fluff* is usually produced after different comminution-classification stages. The final classification is usually carried out by cycloning or venting (air suction or blowing systems), in order to separate the light material. A good separation could contribute to reducing the waste disposal and environmental pollution, and increasing the energy recovery through pure sorted polymer re-use. Furthermore, the possibility to utilize finer *fluff* fractions to produce energy could dramatically contribute to increase the full reutilization of such products. To reach this goal the quantity and the quality of the metal contaminants have to be strongly controlled in order to not prejudice the quality of the final fluff based fuel. The need to develop both efficient selection and control strategies to obtain contaminant-free, or almost free, fine *fluff* products thus assumes a fundamental role in all the processing and control stages of the recycling chain.

Results show the HSI proposed architecture allows identification of all *fluff* constituents and polluting materials (Fig. 6). The discriminating power is high in the region between (400–1,000) nm, this is a quite important result being thus possible to utilize, for *fluff* inspection, a single device (ImSpector™ V10E) sorting/control unit, with considerable costs reduction. From this perspective, the HSI approach can be profitably applied utilizing “simple” band intensity comparison, at specific wavelengths, among unknown *fluff* particles and a reference library of spectra, previously



Recycling Technologies. Figure 5

Representation of the HSI data set, as acquired (**a**), of different *cullets* resulting from a recycling process. The fragment on the right end side of the image is ceramic glass. (**b**) Corresponding false color image embedding the results of all the three score plots, (image of scores on PC1: 97.46, image of scores on PC2: 1.82 and image of scores on PC3: 0.43) related to PC1, PC2, and PC3 components as resulting from the application of the PCA. Contaminant (ceramic glass) can be easily identified



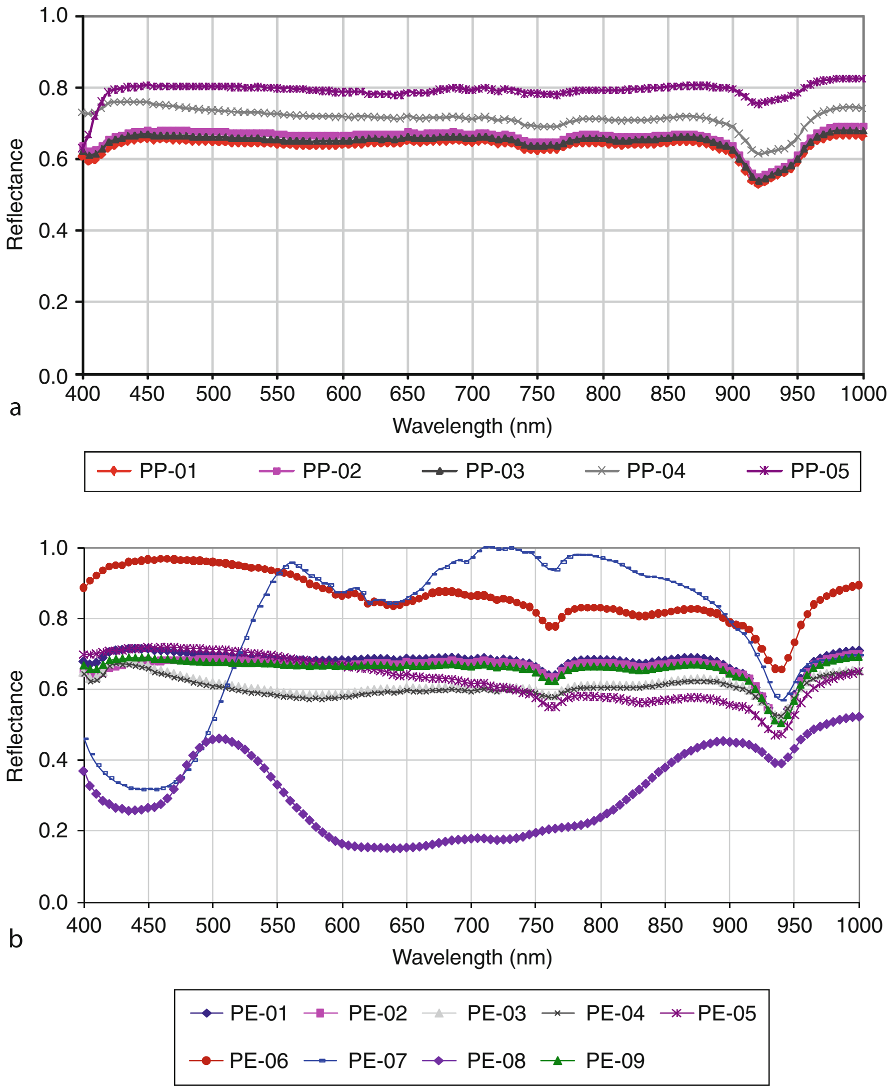
Recycling Technologies. Figure 6

Spectral plots in the VIS–NIR field (400–1,000 nm) of *fluff* materials. The HSI approach clearly allows discrimination between the different light materials constituting *fluff*

built, embedding the different waste materials spectra constituting the light fractions to investigate. The approach is particularly flexible to use because *fluff* of different origin and/or composition can be sorted/controlled adopting the same logic, but working on different specifically spectra library, built according to new materials characteristics to detect.

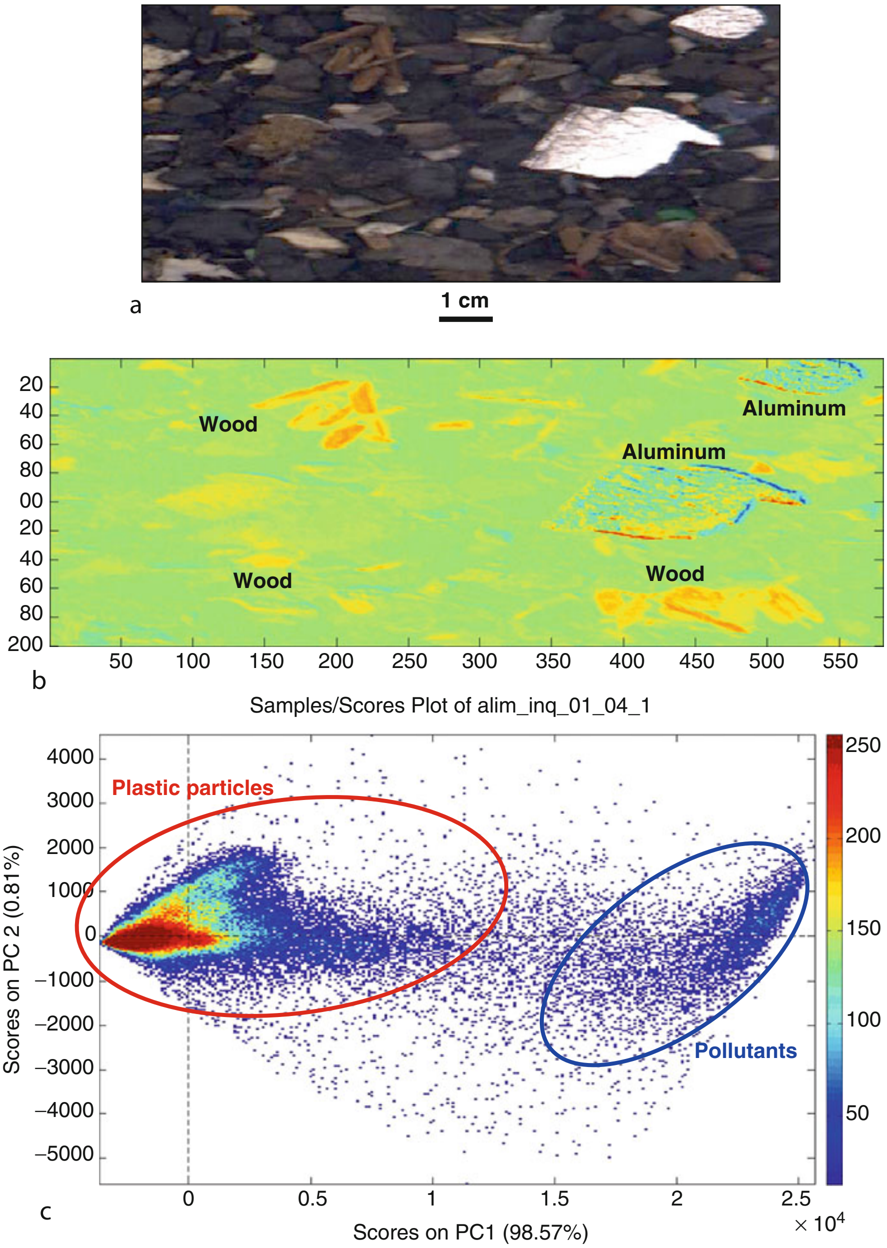
Case Study No. 3: Polyolefins Recognition-Separation. Issues: Polyolefin recovery from complex waste streams is a challenging issue that has not yet been profitably solved. Furthermore, polypropylene (PP), high density polyethylene (HDPE), and low density polyethylene (LDPE) together are both difficult to separate and chemically incompatible. In order to produce high-purity granulates from these concentrates, of a quality comparable to materials produced from post-industrial waste, the mixture must be sorted very accurately, and in order to be economically and ecologically sound, most of the polyolefins should end up in a useful product. Such accurate and efficient separations exist, but they involve multiple separations. They are therefore expensive, difficult to control, and often do not allow the production of good concentrates. The possibility to develop efficient and low-cost recognition logics to control the process and certify the products thus represents a key issue in polyolefin recovery.

State of the art: Innovative processing plant layouts have been realized to process complex plastic-based wastes by shredding and sink-floating to produce polyolefin concentrates of varying quality. Analysis of such concentrates generally shows a mixture of polyolefins, rubbers, foams, fibers, and wood, next to varying amounts of materials heavier than water. Currently available separation techniques, based on the difference in flotation proprieties in water, can be used to separate lighter types of plastic such as PP, HDPE, and LDPE from the heavier types such as polyethylene terephthalate (PET) and polyvinyl chloride (PVC). A known method is to separate the mixtures into five fractions using separation media with densities



Recycling Technologies. Figure 7
Spectral plots in the VIS–NIR field (400–1,000 nm) of polyolefins: PP (**a**) and PE (**b**)

of 880, 920–930, 940, and 970 kg/m^3. Such a procedure will create high-purity PP and HDPE products, whereas foams, most of the wood and rubbers, LDPE, filled PP, and residual heavy materials will end up in relatively small residue fractions. Certain expensive liquids have been specially designed to separate at one of the target densities for polyolefin recycling. Other technologies, such as electrostatic separation and thermal adhesion, have been able to create only a single relatively pure product.



Recycling Technologies. Figure 8

Image on scores on PC2:0.81 (**b**) of the HSI data set of (**a**). Samples/Scores Plot PC1-PC2 clearly outlines the correspondence existing among typology of contaminants and their “mapping” inside the plot (**c**)

Analyses show PP and PE recognition is strongly influenced, in the visible region, by the color of the samples. Based on the analysis of HSI spectra, it is evident the possibility to define specific parameters useful for recognition of the two polyolefins, as, for example, the slope of the spectrum in a selected wavelength range or a band ratio among two different wavelengths. Analyses show PP and PE spectra present significant differences in regions around 750 nm and (900–950) nm (Fig. 7). It is important to note that the best and most precise recognition logic, especially for polluting particles detection and their topological assessment inside plastic waste feed, should be based on more sophisticated and complex statistical analyses, such as (PCA) (Fig. 8), Partial Least Square (PLS), Neural Network (NN), etc., that usually require long computation time. Considering that in most industrial applications the fast response of the detecting/sorting device is one of the main constraints, as, for example, when particles are moving on a conveyor belt and they are sorted *on-line*, the adoption of simplified logics, working properly, is preferred. This latter approach (multiple bands intensity comparison at specific wavelengths) can be profitably applied for recovered polyolefins quality control.

The examples clearly outline the importance of HSI in recycling technologies as an innovative, flexible, and low-cost tool that, combining imaging and reflectance spectroscopy, can profitably allow performance of both waste feed and recovered products characterization, control, and certification. HSI-based technology through the detection of the spectral signature of waste and waste-derived products, of different nature and composition, allows to extract and quantify those physical-chemical attributes influencing their characteristics and behavior. Results demonstrated as the proposed technique, and the related recognition logics, will have a greater impact on the development of recycling technologies finalized to implement *on-line* innovative sorting strategies, as well as new control procedures. The possibility to reach a primary goal in recycling, that is a full control, at a low cost, of the quality of the different flow streams inside the plant, according to the different processing stages, can strongly contribute to develop *innovative-inside-processing* products certification.

Bibliography

Primary Literature

1. Kahn CH (1979) The art and thought of Heraclitus. Cambridge University Press, Cambridge
2. Mamedbeii GD (1959) Muhammed Nasir al-Din al-Tusi on the theory of parallel lines and the theory of ratios. (Azerbaijani), Izdat. Akad. Nauk Azerbaijzansk. SSR (Baku)
3. Dumas M (1955) Lavoisier, théoricien et expérimentateur. Presses Universitaires de France, Paris
4. Heijungs R, Huppes G, Guinée JB (2010) Life cycle assessment and sustainability analysis of products, materials and technologies. Toward a scientific framework for sustainability life cycle analysis. Polym Degrad Stab 95(3):422–428
5. Bartl A, Hackl A, Mihalyi B, Wistuba M, Marini I (2005) Recycling of fibre materials. Process Safety Environ Protect 8(B4):351–358
6. Confederation of European Paper Industries (2009) CEPI sustainability report. Brussels, Belgium
7. Ochoa JAG (2008) Feasibility of recycling pulp and paper mill sludge in the paper and board industries. Resour Conserv Recycl 52(7):965–972
8. Wiegand PS, Unwin JP (1994) Alternative management of pulp and paper industry solid wastes. Tappi J 77:91–97
9. Wolfer EP, Venkat WB, Maroju BV, Martiny A (1997) Method for recovering fiber from effluent streams. U.S. Patent 5, pp 593–542
10. Saint Amand FJ (1999) Hydrodynamics of deinking flotation. Int J Miner Process 56:277–316
11. Tandy S, Healey JR, Nason MA, Williamson JC, Jones DL (2009) Heavy metal fractionation during the co-composting of bio-solids, deinking paper fibre and green waste. Bioresour Technol 100(18):4220–4226
12. Moo-Young HK Jr, Zimmie TF (1997) Waste minimization and re-use of paper sludges in landfill covers: a case study. Waste Manag Res 15(6):593–605
13. Werther J, Ogada T (1999) Sewage sludge combustion. Prog Ener Combust Sci 25(1):55–116
14. Safeglass (Europe) Limited, Nasmyth Building, Nasmyth Avenue. East Kilbride. UK G75 0Q. http://www.breakglass.org/Glass_making.html
15. Höland W, Beall G (2002) Glass–ceramic technology. The American Ceramic Society, Westerville, p 372
16. Pannhorst W (1997) Glass ceramics: state of the art. J Non-Cryst Solids 219:198–204
17. Rem PC (1999) Eddy current separation. Delft University of Technology, Delft
18. Bonifazi G, D'Addetta A, Massacci P (2002) Classification by neural net of a particle stream in an eddy-current drum separator. Int J Part Part Syst Charact 19:96–102
19. Jong TPR, de Dalmijn WL (2002) X-ray transmission imaging for process optimisation of solid resources. In: R02, 6th World congress on integrated resources management, Geneva, Switzerland, CD-Paper 173

20. Bonifazi G (2000) Imaging based sorting logic in solid waste recycling. In: Proceedings of the 16th international conference on solid waste technology and management, vol 6, Philadelphia, USA, pp 14–26
21. Bonifazi G, Massacci P (2000) Cullets (glass fragments) quality control by artificial vision: a textural based approach. In: 4th World Congress R00 – Recovery, recycling, re-integration, Toronto, Canada, CD-Paper 31, pp 723–728
22. Bonifazi G, Massacci P (1998) Cullets (glass fragments) quality control by artificial vision: a color based approach. In: Proceedings of international conference on quality control by artificial vision, Takamatsu, Japan, pp 94–99
23. Serranti S, Bonifazi G, Pohl R (2006) Spectral cullet classification in the mid-infrared field for ceramic glass contaminants detection. Int J Waste Manag Res 24:48–59
24. Bonifazi G, Serranti S (2006) Imaging spectroscopy based strategies for ceramic glass contaminants removal in glass recycling. Int J Waste Manag 26:627–639
25. Cramb AW (1996) A short history of metals. Dept. of Materials Science and Engineering. Carnegie Mellon University. http://neon.mems.cmu.edu/cramb/Processing/history.html
26. Gascoigne B (2001) History of metallurgy. HistoryWorld. Ongoing. http://www.historyworld.net/wrldhis/PlainTextHistories.asp?historyid=ab16
27. Alter H (1977) Magnetic separation – Recovery of salable iron and steel from municipal solid waste. Environmental Protection Agency, Cincinnati
28. Shapiro M, Galperin V (2005) Air classification of solid particles: a review. Chem Engin Process 44:279–285
29. Wills BA (1997) Mineral processing technology, 6th edn. Butterworth-Heinmann, Boston, pp 232
30. Bradley D (1965) The hydrocyclone. Pergamon, New York
31. Takoungsakdakun T, Pongstabodee S (2007) Separation of mixed post-consumer PET–POM–PVC plastic waste using selective flotation. Sep PurifTechnol 54:248–252
32. Buchan R, Yarar B (1995) Recovering plastics for recycling by mineral processing techniques. J Miner Met Mater Soc 47:52–55
33. Drelich J, Kim JH, Payne T, Miller JD, Kobler RW (1999) Purification of polyethylene terephthalate from polyvinyl chloride by froth flotation for the plastics (soft-drink bottle) recycling industry. Sep Purif Technol 15:9–17
34. Kang H, Schoenung JM (2005) Electronic waste recycling: a review of US infrastructure and technology options. Resour Conserv Recycl 45(4):368–400
35. Veit HM, Pereira C, Bernardes AM (2002) Using mechanical processing in recycling printed wiring board. J Miner Met Mater Soc 54(6):45–47
36. Askvik KM, Hetlesæther S, Sjobölm J, Stenius S (2001) Properties of the lignosulfonate–surfactant complex phase. Colloids Surf A Physicochem Eng Aspects 182:178–189
37. Singh BP (1998) Wetting mechanism in the flotation separation of plastics. Filtration Sep 35:525–527
38. Shen H, Pugh RJ, Forssberg E (2002) Floatability, selectivity and flotation separation of plastics by using a surfactant. Colloids Surf A Physicochem Eng Aspects 196:63–70
39. Fraunholcz N (2004) Separation of waste plastics by froth flotation – a review. Part I, Miner Engin 17:261–268
40. Andrady AL (2003) Plastics and the environment. Wiley, Hoboken, pp 792
41. Al-Salem SM, Lettieri P, Baeyens J (2009) Thermal pyrolysis of high density polyethylene (HDPE). In: Proceedings of the 9th European gasification conference: clean energy and chemicals, Düsseldorf, Germany
42. Scheirs J (1998) Polymer recycling: science, technology and application, 1st edn. Wiley-Blackwell, New York
43. Dirks E (1996) Energy recovery from plastic waste in waste incineration plants. In: Brandrup J, Bittner M, Menges G, Michaeli W (eds) Recycling and recovery of plastics, 1st edn. Hanser, Munich, pp 746–769
44. Zia KM, Bhatti HN, Bhatti IA (2007) Methods for polyurethane and polyurethane composites, recycling and recovery: a review. React Funct Polym 67(8):675–692
45. Hawn K (2001) An overview of commercial recycling technologies and textile applications for the products. In: 6th annual conference on recycling of polymer, textile and carpet waste, Dalton, USA
46. Cupit MJ (1996) Opportunities and barriers to textile recycling, AEA Technology, Report 0113, Oxfordshire, UK
47. Passas R, Voillot C, Tarrajat G, Caucal G, Khelifi B, Tourtollet G (2001) Morfi as a novel technology for morphological analysis of fibers. Recents Progres en Genie des Procedes 15: 259–264
48. Bartl A, Mihalyi B, Marini I (2004) Applications of renewable fibrous materials. Chem Biochem Engin 18:21–28
49. Wang Y (1995) Reuse of carpet industrial waste for concrete reinforcement. In: RILEM Proceeding (Disposal and recycling of organic and polymeric construction materials), vol 27, London, pp 297–306
50. Wang Y (1997) Properties of concrete reinforced with recycled carpet waste fibers. In: Proceedings of International symposium on brittle matrix composites 5, Warsaw, pp 179–186
51. Wang Y (1999a) Ecotextile'98: sustainable development. In: Proceedings of the Conference, Bolton, pp 165–171
52. Wang Y (1999) Utilization of recycled carpet waste fibers for reinforcement of concrete and soil. Polym Plast Technol Engin 38:533–546
53. Wang Y (2002) Recycling of automotive fibers. In: Proceedings of Joint INDA-TAPPI Conference, Atlanta, pp 160–167
54. Bohnhoff A, Petershans J (2002) De-centralised technology for the sorting of textile floor coverings. In: 7th annual conference on recycling of polymer, textile and carpet waste, Dalton, USA
55. Strzelecki C (2004) Modern solutions for shredding, grinding and re-pelletizing post-industrial fiber, nonwovens and carpet scrap. In: Annual conference on recycling of polymer, textile and carpet waste, Dalton, USA
56. Bacon FC, Holland WR, Holland LH (1998) Method and machine for recycling discarded carpets. U. S. Patent 5, 704, 104
57. Howe MA, White SH, Locklear SG (2001) Method and apparatus for reclaiming carpet components. US Patent 6, 182, 913

58. Herlihy J (1997) Recycling in the carpet industry. Carpet and Rug Industry, pp 17–25
59. Kasserra P (1998) Recycling of polyamide 6.6 and 6. In: Prasad PN et al (eds) Science and technology of polymers and advanced materials. Plenum, New York, pp 629–635
60. Hagguist JAE, Hume RM (1993) Carpet reclaimer, U.S. Patent 5, 230, 473
61. Schut JH (1995) Big plans for carpet. Plast World 53:25
62. Booij M, Hendrix JAJ, Frentzen YH (1997) Process for recycling polyamide-containing carpet waste, European Patent 759, 456
63. Frentzen YH, Thijert MP, Zwart RL (1997) Process for the recovery of caprolactam from waste containing nylon by extraction with alkyl phenol, World Patent 97, 03, 04
64. Sarian AK, Handerman AA, Jones S, Davis EA, Adbye A (1998) Recovery of polyamide from composite articles, U.S. Patent 5, 849, 80
65. Sikorski ME (1993) Recycling of polymeric materials from carpets and other multi-component structures by means of supercritical fluid extraction, U.S. Patent 5, 233, 021
66. Griffith AT, Park Y, Roberts CB (1999) Separation and recovery of nylon form carpet waste using a supercritical fluid antisolvent technique. Polym Plast Technol Engin 38(3):411–432
67. Honeywell Nylon Inc (2005) http://www.infinitynylon.com
68. Elam CC, Evan RJ, Czernik S (1997) An Integrated approach to the recovery of fuels and chemicals from mixed waste carpets through thermocatalytic processing, Preprint papers - American Chemical Society. Div Fuel Chem 42(4):993–997
69. Bajaj P, Sharma ND (1997) In: Gupta VB, Kothari VK (eds) Reuse of polymer and fibre waste in manufactured fibre technology. Chapman & Hall, New York, pp 615
70. Brown T (2001) Infinity nylon - a never-ending cycle of renewal, 6th annual conference on Recycling of polymer, Textile and Carpet Waste, Dalton, GE, USA. http://hdl.handle.net/1853/10385
71. Schut JH (1993) A recycling first: carpets! Plast Technol :22–25
72. Young D, Chlystek S, Malloy R, Rios I (1998) Recycling of carpet scrap, U.S. Patent 5,852,115
73. Hagberg CG, Dickerson JL (1997) Recycling nylon carpet via reactive extrusion. Plast Engin 53:41–43
74. Datta RJ, Polk MB, Kumar S (1995) Reactive compatibilization of polypropylene and nylon. Polym Plast Technol Engin 34(4):551–560
75. Dagli SS, Xanthos M, Biesenberger JA (1992) Blends of nylon 6 and polypropylene with potential applications in recycling, effects of reactive extrusion variables on blend characteristics. ACS Symp Ser 513:241–257
76. David DJ, Dickerson JL, Sincock TF (1994) Thermoplastic composition and method for producing thermoplastic composition by melt blending carpet, U. S. Patent 5, 294, 384
77. Muzzy J, Wang Y, Hagberg C, Patel P, Jin K, Samanta S, Bryson L, Shaw B (2004) Long fiber reinforced post-consumer carpet. In: ANTEC 2004, Annual Technical Conference of the society of plastics engineers, Chicago, USA
78. Jähne B (1993) Digital image processing; concepts, algorithms, and scientific applications, 2nd edn. Springer, Berlin
79. de Kattentidt HUR, Jong TPR, Dalmijn WL (2003) Multi-sensor identification and sorting of bulk solids. Control Engin Pract 11:4147
80. Bearmann GH, Levenson RM, Cabib D (eds) (2002) Spectral imaging: basic principles and prospective applications. Kluwer, Dordrecht
81. Leitner R, Mairer H, Kercek A (2003) Real-time classification of polymers with NIR spectral imaging and blob analysis. Real-Time Imag 9:245–251
82. American Wood Preservers' Association (AWPA) (1999) American wood preservers' association book of standards. American Wood Preservers' Association, Grandbury
83. Blassino M, Solo-Gabriele HM, Townsend T (2002) Pilot scale evaluation of sorting technologies for CCA treated wood waste. Waste Manag Res 20:290–301
84. Kormienko M (1999) Sorting technologies for CCA-treated wood waste. Master of Science Thesis, University of Miami, Coral Gables, USA
85. Solo-Gabriele H, Townsend T, Kormienko M, Stook K, Gary K, Tolaymat T (2000) Alternative chemicals and improved disposal-end management practices for CCA-treated wood. Final Technical Report #00-03. Florida center for solid and hazardous waste management, Gainesville, USA
86. de Jong TPR, Dalmijn WL (2002) X-ray transmission imaging for process optimisation of solid resources. In: Proceedings of R'02 congress, Geneva, pp 1–6
87. Hahn DW, Flower WL, Hencken KR (1997) Discrete particle detection and metal emissions monitoring using laser-induced breakdown spectroscopy. Appl Spectrosc 51: 1836–1844
88. Hahn DW (1998) Laser-induced breakdown spectroscopy for sizing and elemental analysis of discrete aerosol particles. Appl Phys Lett 72:2960–2962
89. Radziemski LJ, Cremers DA (1989) Laser-induced plasmas and applications. Marcel Dekker, New York
90. de Mesina MB, Jong TPR, Dalmijn WL (2007) Automatic sorting of scrap metals with a combined electromagnetic and dual energy X-ray transmission sensor. Int J Miner Process 82:222–232
91. Goetz AFH, Vane G, Solomon TE, Rock BN (1985) Imaging spectrometry for earth remote sensing. Science 228: 1147–1153
92. Hege E, O'Connell D, Johnson W, Basty S, Dereniak E (2003) Hyperspectral imaging for astronomy and space surveillance. Proceedings of the SPIE 5159:380–391
93. Wood KS, Gulian AM, Fritz GG, Van Vechten D (2002) A QVD detector for focal plane hyperspectral imaging in astronomy. Bull Am Astron Soc 34:1241
94. Monteiro S, Minekawa Y, Kosugi Y, Akazawa T, Oda K (2007) Prediction of sweetness and amino acid content in soybean crops from hyperspectral imagery. ISPRS J Photogram Remote Sens 62(1):2–12
95. Smail V, Fritz A, Wetzel D (2006) Chemical imaging of intact seeds with NIR focal plane array assists plant breeding. Vibrational Spectroscopy 42(2):215–221

96. Lyon RC, Lester DS, Lewis EN, Lee E, Yu LX, Jefferson EH (2002) Near-infrared spectral imaging for quality assurance of pharmaceutical products: analysis of tablets to assess powder blend homogeneity. AAPS Pharm Sci Tech 3(3):17
97. Rodionova O, Houmøller L, Pomerantsev A, Geladi P, Burger J, Dorofeyev V (2005) NIR spectrometry for counterfeit drug detection: a feasibility study. Anal Chim Acta 549(1–2):151–158
98. Roggo Y, Edmond A, Chalus P, Ulmschneider M (2005) Infrared hyperspectral imaging for qualitative analysis of pharmaceutical solid forms. Anal Chim Acta 535(1–2):79–87
99. Ferris D, Lawhead R, Dickman E, Holtzapple N, Miller J, Grogan S (2001) Multimodal hyperspectral imaging for the non invasive diagnosis of cervical neoplasia. J Low Genit Tract Dis 5(2):65–72
100. Kellicut D, Weiswasser J, Arora S, Freeman J, Lew R, Shuman C (2004) Emerging technology: hyperspectral imaging. Perspect Vasc Surg Endovasc Ther 16(1):53–57
101. Zheng G, Chen Y, Intes X, Chance B, Glickson JD (2004) Contrast-enhanced near-infrared (NIR) optical imaging for subsurface cancer detection. J Porphyrins Phthalocyanines 8(9):1106–1117
102. Serranti S, Bonifazi G (2009) Hyperspectral imaging detection architectures for polyethilene (PE) and polypropylene (PP) identification inside plastic waste streams. In: Proceedings of waste-to-resources, III International symposium MBT&MRF. Hanover, Germany, pp 463–474
103. Serranti S, Bonifazi G, Bonoli A, Dall'Ara A (2009) Composting products quality assessment and monitoring by hyperspectral imaging based logics. In: Proceedings of waste-to-resources, III International symposium MBT&MRF. Hanover, Germany, pp 584–597
104. W2Plastics (2008) Collaborative Project 212782 - FP7-ENV-2007-1: magnetic sorting and ultrasound sensor technologies for production of high purity secondary polyolefins from waste
105. HYSPIMGLASS (2002) CRAFT Programme: CRAF-1999-71817: development of a Novel and high speed spectral imaging system to detect glass-like contaminants in the recyclable, cost-effectively increasing glass recycling and avoiding landfilling
106. SSOM (2008) Spectral scanner operative manual (Version 2.0). DV Optics S.r.l., Italy http://www.dvoptic.com/index.html
107. Bonifazi G (2000) Imaging based sorting logic in solid waste recycling. In: The Sixteenth international conference on solid waste technology and management – session 6A: recycling and source reduction. Philadelphia, USA, pp 6.14–26
108. Geladi P, Isaksson H, Lindqvist L, Wold S, Esbensen K (1989) Principal components analysis of multivariate images. Chemometr Intell Lab Syst 5(3):209–220

Books and Reviews

An good paper presenting an excellent review of recent progress in the recycling and recovery of P*lastic Solid Waste* (PSW), with particular emphasis "*on waste generated from polyolefinic sources, which makes up a great percentage of our daily single-life cycle plastic products*" is: Al-Salem SM, Lettieri P, Baeyens J (2009) Recycling and recovery routes of plastic solid waste (PSW): a review. Waste Manag 29:2625–2643

Beede DN, Bloom DE (1995) Economics of the generation and management of MSW. NBER Working Papers 5116. National Bureau of Economic Research, Inc, Cambridge, MA

Caputo AC, Pelagagge PM (2001) Waste-to-energy plant for paper industry sludges disposal: technical-economic study. J Hazard Mater 81(3):265–283

Cofie O, Kone D, Rothenberger S, Moser D, Zubruegg C (2009) Co-composting of faecal sludge and organic solid waste for agriculture: process dynamics. Water Res 43(18): 4665–4675

El Haggar S (2007) Sustainable industrial design and waste management: cradle-to-cradle for sustainable development. Academic, St. Louis, pp 424

Galperin V, Shapiro M (1999) Separation of solid particles in a fluidized bed air classifier. Powder Handling Process 11:2

Gesing A, Berry L, Dalton R, Wolanski R (2002) Assuring continued recyclability of automotive aluminium alloys: grouping of wrought alloys by color, X-ray absorption and chemical composition-based sorting. In: Proceedings annual meeting on automotive alloys and aluminium sheet and plate rolling and finishing technology, Seattle, USA

Gesing A, Steward C, Wolanski R, Dalton R, Berry R (2000) Scrap preparation for aluminium alloy sorting. In: Proceedings TMS fall extraction and process metallurgy meeting, Pittsburgh, USA

Hosokawa Micron Group (2011) http://www.hmicronpowder.com/application/classification

Huth-Fehre Th, van den Broek W (1995) NIR-Remote sensing and artificial neural networks for rapid identification of post consumer plastics. J Mol Struct 348:143–146

Johansson JE (2007) Plastics – the compelling facts and figures. 6th IdentiPlast Biennial Conference on the Recycling and Recovery of Plastics. Brussels, Belgium

Kunii D, Levenspiel O (1991) Fluidization engineering, 2nd edn. Butterworth, Heinmann, pp 233

Marques GA, Tenorio JAS (2000) Use of froth flotation to separate PVC/PET mixtures. Waste Manag 20:265–269

Méndez A, Fidalgo JM, Guerrero F, Gascó G (2009) Characterization and pyrolysis behaviour of different paper mill waste materials. J Anal Appl Pyrolysis 86(1):66–73

Oshitani J, Kiyoshima K, Tanaka Z (2003) Continuous dry material separation from automobile shredder residue. Kagaku Kogaku Ronbunshu 29:8–14

Pascoe RD (2005) The use of selective depressants for the separation of ABS and HIPS by froth flotation. Miner Eng 18:233–237

Sekito T, Matsuto T, Tanaka N (2006) Application of a gas–solid fluidized bed separator for shredded municipal bulky solid waste separation. Waste Manag 26:1422–1429

Sekito T, Tanaka N, Matsuto T (2006) Batch separation of shredded bulky waste by gas–solid fluidized bed at laboratory scale. Waste Manag 26:1246–1252

Svoboda J (2004) Magnetic techniques for the treatment of materials, Kluver Academic Publisher, New York, USA. pp. 656.

http://www.springer.com/earth+sciences+and+geography/book/978-1-4020-2038-4

Van Nieuwenhuijzen A, Van der Graaf J (2010) Handbook on particle separation processes. IWA, London, UK, pp 400

World Bank (2007) Environmental, health, and safety guidelines for pulp and paper mills. Draft technical document. Environment and Social Development Department, International Finance Corporation, Washington, DC

Yoshida M, Oshitani J, Kaname K, Gotoh K (2006) Fluidized bed medium separation (FBMS) of Cl-containing plastics in home electric appliance shredder residue. Kagaku Kogaku Ronbunshu 32:115–121

Regenerative Braking

Yimin Gao
Department of Electrical and Computer Engineering,
Texas A&M University, College Station, TX, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Braking Energy
Hybrid Brake System
Future Directions
Bibliography

Glossary

Antilock braking system (ABS) Safety system that prevents the wheels on a vehicle from locking up while braking.

Brake-by-wire Technology to replace the traditional mechanical and hydraulic control systems with electronic control systems using electromechanical actuators and human–machine interfaces.

Brake controller A microchip-based device which receives brake strength command from driver and feedback signals and generates control signals to control the mechanical and electric brake systems.

Braking energy The energy consumed by the brake system of a vehicle.

Hybrid brake system A vehicle brake system that uses both regenerative braking and mechanical fictional brake.

Regenerative braking Braking device of a vehicle which can absorb vehicle braking energy and store the absorbed braking energy into an energy storage and then uses it for later traction.

Definition of the Subject

Regenerative brake system is a newly developed brake system used in electric, hybrid electric, and fuel cell vehicles which can convert part of braking energy into electric energy using an electric motor/generator. For braking safety, the traditional mechanical brake is still required. The electric and mechanical brake systems constitute a hybrid braking system. In the design and control of such a system, the braking force distribution on front and rear wheels has to meet the braking regulation for braking safety, and meantime, a braking force distribution on mechanical and electric brake system design should be able to recover as much braking energy as possible.

Introduction

One of the most important features of electric, hybrid, and fuel cell vehicles is the ability of recapturing part of braking energy, which can save significant amount of fuel, especially for the vehicles driving mostly with stop-and-go pattern.

Brake system of a vehicle is most important for vehicle safety, which has to meet the demands of quickly slowing down the vehicle and maintaining the vehicle stability. In normal driving, braking force requirements widely vary from slightly retarding vehicle to emergency braking, depending on the traffic situation. Heavy braking needs a huge braking force that regenerative brake cannot produce, consequently, a mechanical brake has to be engaged. This forms a hybrid brake system. As in a hybrid propulsion system, there are many options in configurations and control strategies. An advanced hybrid braking system may be a braking-by-wire system with intelligent control.

Braking Energy

Percentage of Braking Energy to Propelling Energy

Braking force has to be applied to a vehicle while slowing down its speed or descending a grade. The kinetic energy ($1/2M\,V^2$) or potential energy ($M\,g\,h$, where, h is the vertical height of the vehicle descent) is dissipated through vehicle brake system. For relative flat roads, most of the braking energy is consumed for slowing down the vehicle speed, especially, driving in urban areas. Thus, discussion in this section focuses on the urban stop-and-go driving pattern. The stop-and-go driving pattern is interpreted by typical driving cycles, such as those of EPA FTP 75 urban, LA 92, US 06, ECE-15, and New York city. Of course, other driving cycles can also be used.

While a vehicle driving on a flat road, the braking force measured on drive wheels for developing a deceleration j (m/s^2) can be expressed as

$$F_b = M\delta j - F_r - F_w, \tag{1}$$

where, M is the vehicle mass in kg, δ is the equivalent mass factor, which is a factor that equivalently converts the rotating inertia into translational inertia, F_r is the rolling resistance, and F_w is the aerodynamic drag of the vehicle. Correspondingly, the braking power can be expressed as

$$P_b = \frac{V}{1000}(M\delta j - F_r - F_w),\ (kW) \tag{2}$$

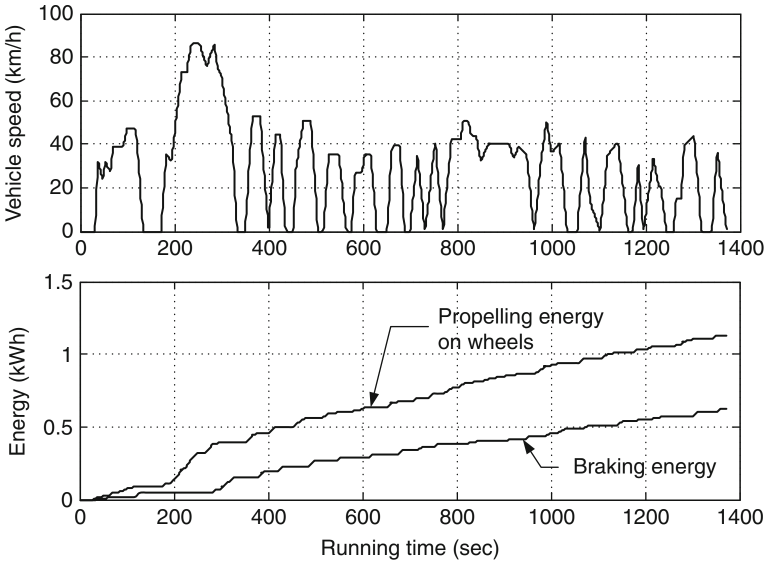
where, V is vehicle speed in m/s. The energy consumed by braking can be obtained as

$$E_b = \int P_b\,dt. \tag{3}$$

Figure 1 shows the propelling and braking energy profiles of 1,500 kg passenger car while driving in EPA FTP 75 urban driving cycle. It can be seen that, propelling energy consumed on the wheels in a whole cycle is about 1.13 kWh, in which, about 0.63 kWh is consumed in braking. The braking energy is about 55.4% of the total propelling energy which comes from the power sources of the vehicle. Table 1 lists the energy scenario of the 1,500 kg passenger car in the driving cycles of FTP 75 urban, LA92, US06, ECE-15, and New York city. The data in Table 1 shows a significant percentage of the energy consumed by braking. It also hints the importance of an effective regenerative brake system in energy saving.

Braking Energy Distribution over Vehicle Speed

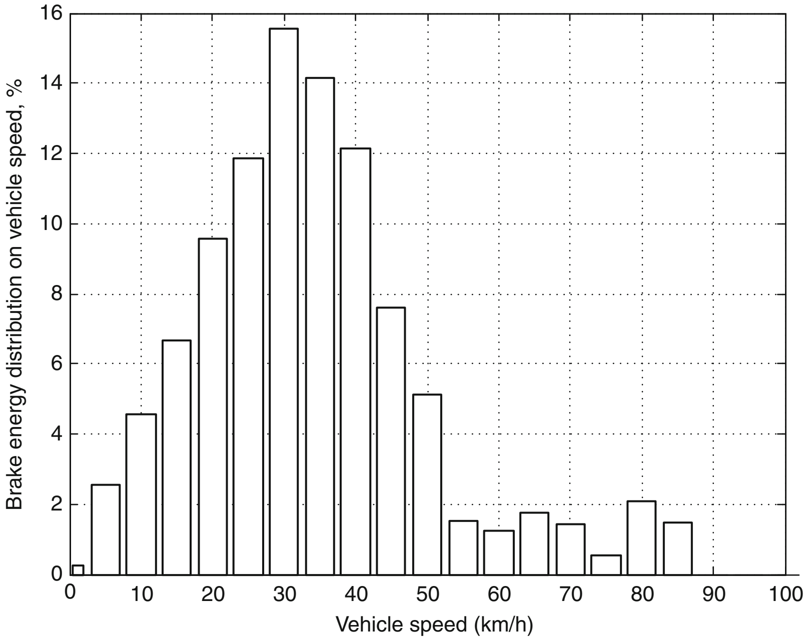
Understanding the braking energy distribution over vehicle speeds is helpful for an effective generative brake system design. For example, in regenerative brake system design, attempts may be made for maximizing regenerative braking capability in the speed range in which most significant braking energy appears. In other speed ranges with less braking energy, other braking performance may be the major concern without significant scarification of energy recovery.



Regenerative Braking. Figure 1
Significant percentage of propelling energy consumed in braking [1]

Regenerative Braking. Table 1 Braking energy scenario of 1,500 kg passenger car in FTP 75 urban drive cycle [1]

	FTP75 urban	LA92	US06	ECE-15	New York city
Maximum speed (km/h)	86.4	107.2	128.5	120	44.6
Average speed (km/h)	27.9	39.4	77.5	49.9	12.2
Traveling distance per cycle (km)	10.6	15.7	12.8	7.95	1.9
Propelling energy (kWh)					
Per cycle	1.129	2.356	2.266	0.969	0.296
Per km	0.106	0.15	0.177	0.122	0.156
Braking energy (kWh)					
Per cycle	0.625	1.337	0.923	0.330	0.243
Per km	0.059	0.087	0.072	0.042	0.127
Percentage of braking energy to propelling energy	55.4	58.0	40.7	34.1	81.9



Regenerative Braking. Figure 2
Braking energy density over vehicle speeds in FTP75 urban driving cycle

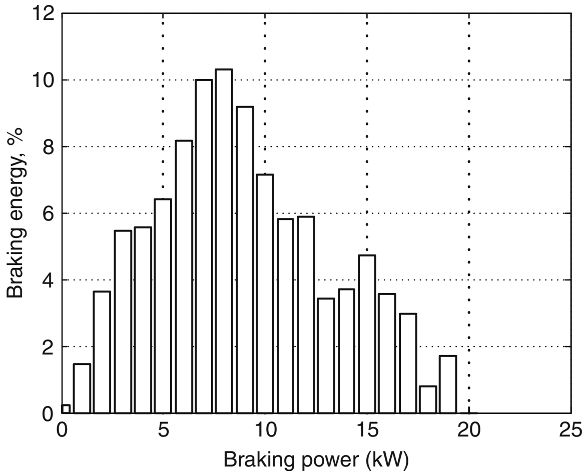
In calculation of the braking energy distribution over vehicle speeds, the whole speed range is divided into a serial of speed interval, $\varDelta V_j$ ($j = 1, 2, \ldots, n$). The braking energy in the speed interval $\varDelta V_j$ can be obtained by integrating the powers falling into the speed interval $\varDelta V_j$, with respect to time, as

$$\varDelta E_b^j = \int\limits_{\varDelta V_j} P_b^j dt, \tag{4}$$

where $\varDelta E_b^j$ and $P_b^j$are the braking energy and braking power in the speed interval of $\varDelta V_j$. Figure 2 shows

Regenerative Braking. Table 2 Braking energy scenario of 1,500 kg passenger car in FTP 75 urban drive cycle [1]

	FTP 75 urban	LA92	US06	ECE-15	NY city
Braking energy @ speeds <15 km/h (%)	10.93	5.51	3.27	4.25	21.32



Regenerative Braking. Figure 3
Braking energy distribution over braking power in FTP75 urban driving cycle

Regenerative Braking. Table 3 Braking range in which 85% of braking energy falls dissipated in typical urban driving cycles [1]

	FTP 75 urban	LA92	US06	ECE-15	NY city
Power range in which 85% of total energy falls (kW)	<14.4	<44.5	<46.5	<33.5	<18.5

a vehicle braking energy distribution diagram over vehicle speeds in FTP 75 urban driving cycle. It can be seen that most of the braking energy are consumed in the speed range of 15–50 km/h and very small amount falls into the speed range of less than 15 km/h and larger than 50 km/h. It has been known that electric motor has a poor performance for generation in low speed due to its low back electromagnetic force (EMF). Thus, in regenerative brake system design, regenerative braking may be cancelled and only mechanical braking is applied at low speed, which will not sacrifice much braking energy recovery. Table 2 shows the braking energy consumed in the speed range of less then 15 km/h in typical driving cycle.

Braking Energy Distribution over Braking Power

Braking energy distribution over braking power can supply useful information for the power capacity design of an electric, especially, for mild hybrid vehicle, in which a small electric motor is usually used and for battery design, so that they can recover most of the braking energy.

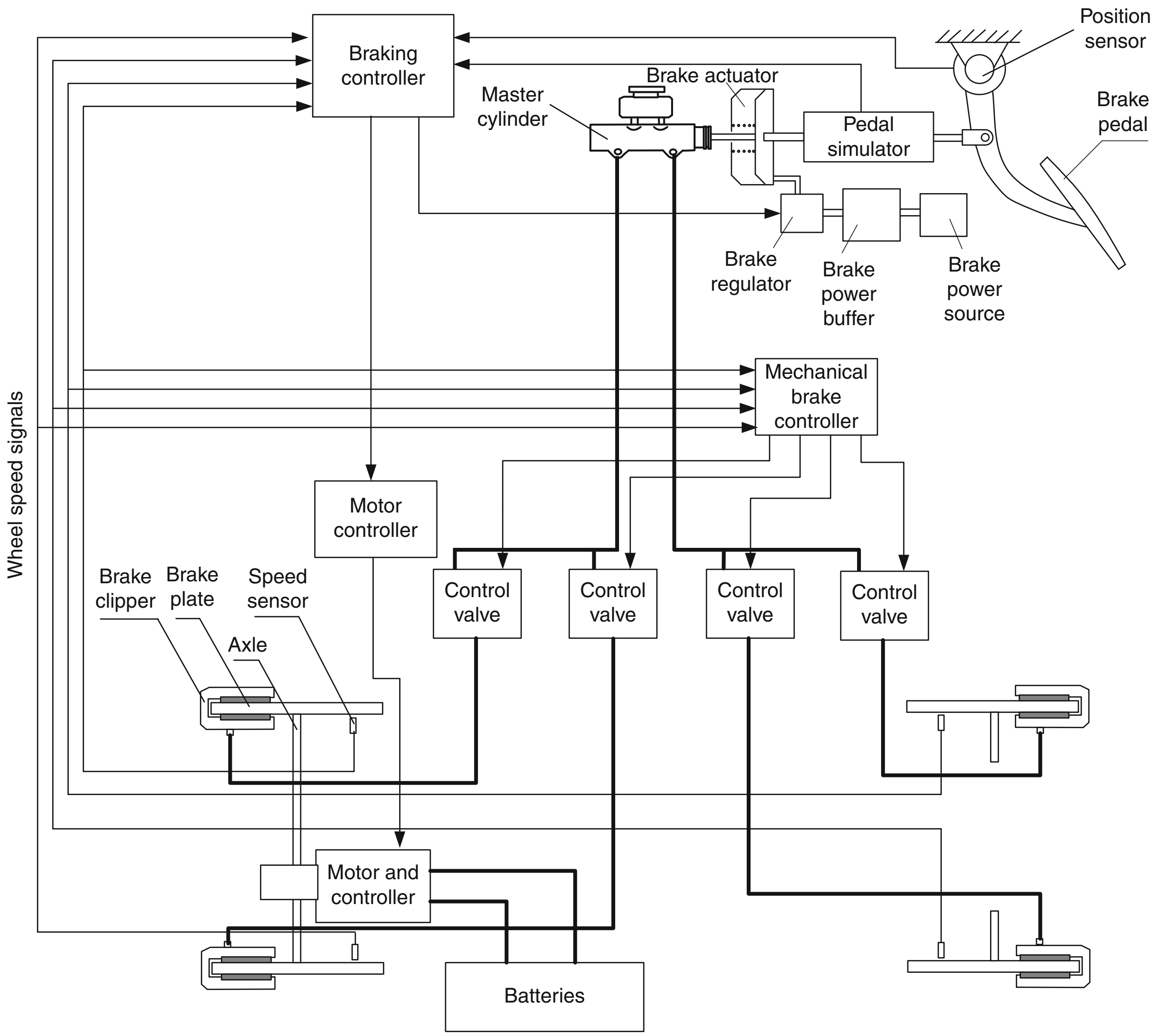
Similar to the braking energy distribution on vehicle speeds, the braking power is divided into a series of power intervals $\varDelta P_j$ and the braking energy in a power interval $\varDelta P_j$ is calculated by

$$\varDelta E_b^j = \int\limits_{\varDelta P_j} P_b^j \, dt, \tag{5}$$

where $\varDelta E_b^j$ and P_b^j are the braking energy and braking power interval of $\varDelta P_j$. Figure 3 shows the braking power distribution over braking powers of a 1,500 kg passenger car, in which 85% of braking energy falls into the power range of less than 15 kW. Table 3 gives the power ranges in which 85% of the braking is consumed in the driving cycles of FTP 75 urban, LA92, US06 ECE-15, and New York city.

Hybrid Brake System

A brake-by-wire system is conceptually illustrated in Fig. 4. A brake pedal simulator is connected to a brake pedal. The brake simulator is a mechanical or hydraulic device that produces a "braking feel" for the driver. Instead of generating a braking force, the brake simulator generates a braking command signal which represents driver's desired braking force. The braking controller controls a brake regulator for producing a correct thrust force for the brake actuator. The brake actuator pushes the master cylinder rod. A brake power buffer may be used for supplying brake thrust force and a brake power source



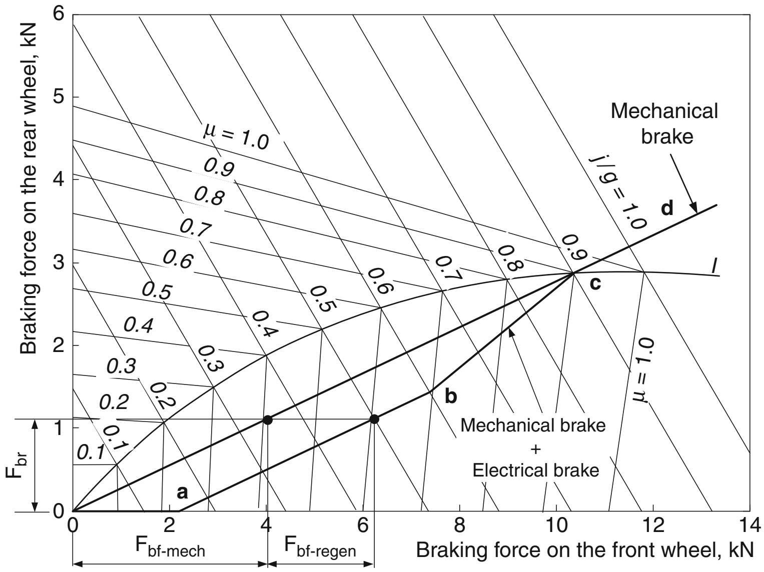
Regenerative Braking. Figure 4
A brake-by-wire hybrid braking system

powers the brake system. This system may be pneumatic, which includes an air solenoid as the brake actuator, air pressure regulator as the brake regulator, a compressed air storage as the brake power buffer, and an air compressor as the brake power source.

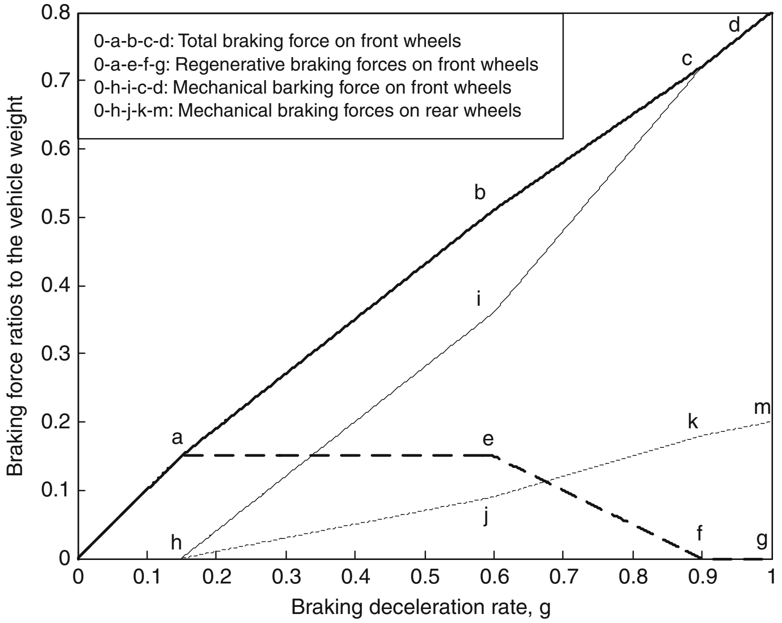
A brake position sensor is also attached to the brake pedal to sense its position. Based on receiving braking command and feedback speed signals, the computing algorithm embedded in the brake controller generates control commands to the electric motor and the pressure regulator. In the case of failure of the electric system, the brake simulator can directly apply its force to the master cylinder to generate a mechanical braking force.

The mechanical brake controller operates very similar to the controller of a conventional anti-lock brake system (ABS), which operates the control values to produce correct mechanical brake torque to each of the wheels to preventing wheels from lockup.

The braking controller functions for correctly allocating total braking force to electric regenerative



Regenerative Braking. Figure 5
Braking forces varying with braking strength



Regenerative Braking. Figure 6
Braking force ratios relative to the vehicle weight versus braking strength [1]

braking and mechanical braking. There are several control strategies for this purpose.

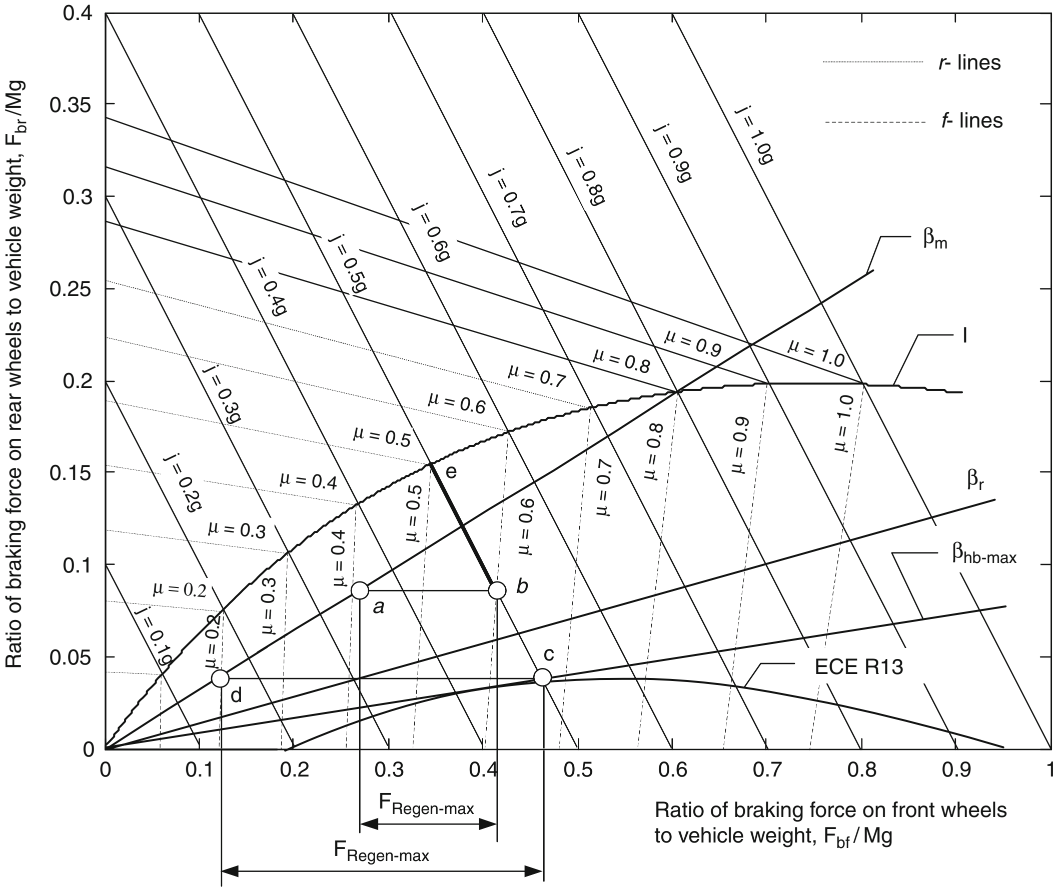
A simple control strategy is shown in Fig. 5, where, front wheels are braked by both regenerative and mechanical brakes, and rear wheels only by mechanical brake. This control strategy is explained as below.

With a slight braking strength, for example, less than 0.15 g, only regenerative braking on front wheels is applied. This operation is simulating the engine retarding effect. Actually, in normal driving, the braking strength falls in this range in most of driving time [1]. Since no mechanical braking is applied to both the front and the rear wheels, the braking torque generated by the electric machine may be commanded by the position of the brake pedal. However, at low speed, the mechanical braking may be needed due to the weak regenerative braking capability. Furthermore, the electric motor should have sufficient torque to fulfill this braking action.

Regenerative Braking. Table 4 The percentage of the total braking energy available for recovering

	FTP 75 urban	LA 92	US 06	New York	ECE-15
Percentage of the total braking energy available for recovering	89.69	82.92	86.55	76.16	95.75

When strong braking is required, both mechanical and regenerative brakes are needed to be applied to front wheels and only mechanical braking is applied to rear wheels. The braking force distribution on the front and rear wheels follows the lines of a–b–c–d and mechanical braking distribution follows the straight line a–d. In the braking strength range of 0.15–0.6 g, the regenerative braking part is a constant as shown in Fig. 6. When the braking strength is greater than 0.6 g, the regenerative braking is gradually reduced to zero at a high braking strength of 0.9 g. At high braking strength, mechanical brake is more reliable. Figure 6



Regenerative Braking. Figure 7
Illustration of braking force distribution on the front wheels (electrical + mechanical) and rear wheels

shows an example of total braking force scenarios. Table 4 shows the calculation results of the percentage of the braking energy that can be recovered by this control strategy in typical urban driving cycles [1].

A control strategy that focuses on maximum regenerative braking may be applied. This braking control strategy follows the rule of allocating the total braking force to front wheels as much as possible within the range stipulated by regulation, for example, ECE R13 regulation. The ECE R13 regulation stipulates a braking force distribution zone, which is in between curves *I* and ECE R13 as shown in Fig. 7. For control simplicity, a straight line, $\beta_{hb\text{-}max}$ may replace the ECE R13 curve.

For certain braking strength commanded by the driver (e.g., 0.5 g) on a road with adhesive coefficient (e.g., 0.6), the braking force distribution can be at any point in the segment of b–e as shown in Fig. 7. Obviously, operating at point b, front wheels obtain its maximum braking force. In this case, the available regenerative braking force on the front wheels is the segment of a-b, depending on the motor torque capacity. The mechanical brake works at point a. If the road coefficient increases to 0.7 and still with braking strength 0.5 g, the braking force distribution can be at any point between points c–e. Obviously, the maximum available braking force on the front wheels is at point c, limited by the ECE R13 regulation. The mechanical brake works at point d and available regenerative braking force is the segment of c–d. This control strategy can be summarized as a rule such that, if the road adhesion and motor torque permit, operate the total braking force distribution on the line $\beta_{hb\text{-}max.}$ Obviously, line $\beta_{hb\text{-}max}$ stands for the extreme case. In practice, a margin would be set, for example, line β_r as shown in Fig. 7.

Future Directions

A fully electronic controlled, brake-by-wire hybrid braking system is still under development. In this system, the electric motor/generator and mechanical braking force on individual wheels can be independently and coordinately controlled. While driving on any road, the road adhesive capacity can be fully used for quickly stopping the vehicle, maintaining the vehicle stability, and meanwhile recovering braking energy as much as possible. The braking system is coordinately controlled with steering system for vehicle driving stability.

Bibliography

Primary Literature

1. Ehsani M, Gao Y, Emadi A (2010) Modern electric, hybrid and fuel cell vehicles-fundamentals, theory and design. CRC, Boca Raton

Books and Reviews

Cikanek SR, Bailey KE (1995) Energy recovery comparison between series and parallel braking system for electric vehicle using various drive cycles, DSC vol 56/DE, 86. Advanced Automotive Technologies, American Society of Mechanical Engineers (ASME), New York, pp 17–31

Gao H, Gao Y, Ehsani M (2001) Design issues of the switched reluctance motor drive for propulsion and regenerative braking in EV and HEV. Proceedings of the SAE 2001 future transportation technology conference, Costa Mesa, CA, paper No. 2001-01-2529

Gao Y, Chu L, Ehsani M (2007) Design and control principle of hybrid braking system for EV, HEV and FCV, 2007 IEEE VPPC

Shu J, Zhang Y, Yin CL (2009) Longitudinal control of hybrid electric buses using traction motor and pneumatic braking system. WSEAS Trans Circuits Syst 8(11):873–882

Regenerative Development and Design

Pamela Mang[1], Bill Reed[2]
[1]Regenesis Group and Story of Place Institute, Santa Fe, NM, USA
[2]Regenesis Group and Story of Place Institute, Arlington, MA, USA

Article Outline

Glossary
Definition of the Subject and Its Importance
Introduction
Regenerative Development and Design–Redefining Sustainability
Overview: Ecological Sustainability and Regenerative Development and Design

Regenerative Approaches to Sustainable Development and Design – Key Framework Premises and Practice Methodologies Overview
Future Directions
Bibliography

Glossary

Biomimicry Sometimes called biomimetic design, an emerging design discipline that looks to nature for sustainable design solutions [1].

Cradle to Cradle® Framework for designing manufacturing processes "powered by renewable energy, in which materials flow in safe, regenerative, closed-loop cycles," and which "identifies three key design principles in the intelligence of natural systems, which can inform human design: *Waste Equals Food*; *Use Current Solar Income*; *Celebrate Diversity*" [2, 3].

Ecoliteracy The ability to understand the natural systems that make life on earth possible, including understanding the principles of organization of ecological communities (i.e., ecosystems) and using those principles for creating sustainable human communities [4, 5].

Ecological sustainability A biocentric school of sustainability thinking that, based on ecology and living systems principles, focuses on "the capacity of ecosystems to maintain their essential functions and processes, and retain their biodiversity in full measure over the long term"; contrasts with technological sustainability based on technical and engineering approaches to sustainability [4].

Ecology The interdisciplinary scientific study of the living conditions of organisms in interaction with each other and with the surroundings, organic, as well as inorganic.

Ecosystem "The interactive system of living things and their nonliving habitat" [6].

Ecosystem concept "A coherent framework for redesigning our landscapes, buildings, cities, and systems of energy, water, food, manufacturing, and waste" through "the effective adaptation to and integration with nature's processes." It has been used more to shape an approach than as a scientific theory [7].

Living systems thinking A thinking technology, using systemic frameworks and developmental processes, for consciously improving the capacity to apply systems thinking to the evolution of human or social living systems [8].

Locational patterns The patterns that depict the distinctive character and potential of a place and provide a dynamic mapping for designing human structures and systems that align with the living systems of a place.

Pattern literacy Being able to read, understand, and generate ("write") appropriate patterns.

Permaculture A contraction of *permanent agriculture* or *permanent culture*, permaculture was developed as a system for designing ecological human habitats and food production systems based on the relationships and processes found in natural ecological communities and the relationships and adaptations of indigenous peoples to their ecosystems [9].

Place The unique, multilayered network of ecosystems within a geographic region that results from the complex interactions through time of the natural ecology (climate, mineral, and other deposits, soil, vegetation, water, and wildlife, etc.) and culture (distinctive customs, expressions of values, economic activities, forms of association, ideas for education, traditions, etc.).

Regenerate American Heritage Dictionary of the English Language and Merriam Webster Dictionary

- To give new life or energy; to revitalize; to bring or come into renewed existence; to impart new and more vigorous life
- To form, construct, or create anew, especially in an improved state; to restore to a better, higher or more worthy state; refreshed or renewed
- To reform spiritually or morally; to improve moral condition; to invest with a new and higher spiritual nature
- To improve a place or system, especially by making it more active or successful

Regenerative design A system of technologies and strategies based on an understanding of the inner working of ecosystems that generates designs to regenerate rather than deplete underlying life support systems and resources within socioecological wholes.

Regenerative development A system of technologies and strategies for generating the patterned whole-system understanding of a place, and developing the strategic systemic thinking capacities and the self-organizing and self-evolving stakeholder engagement/commitment required to ensure regenerative design processes achieve maximum systemic leverage and support.

Restorative design Sometimes called restorative environmental design, a design system that combines returning "polluted, degraded or damaged sites back to a state of acceptable health through human intervention" [10] with biophiliac designs that reconnect people to nature.

Source to sink Simple linear flows from resource sources (farms, mines, forests, watershed, oilfields, etc.) to sinks (air, water, land) that deplete global sources and overload/pollute global sinks [11].

Systems thinking A framework for seeing interrelationships rather than things and for seeing patterns of change rather than static "snapshots." It addresses phenomena in terms of wholeness rather than in terms of parts [5].

Definition of the Subject and Its Importance

The emerging field of regenerative development and design marks a significant evolution in the concept and application of sustainability. Practices in sustainable or green design have focused primarily on minimizing damage to the environment and human health, and using resources more efficiently, in effect, slowing down the degradation of earth's natural systems. Advocates of a regenerative approach to the built environment believe that a much more deeply integrated, whole-systems approach to the design and construction of buildings and human settlements (and nearly all other human activities) is needed. Regenerative approaches seek not only to reverse the degeneration of the earth's natural systems but also to design human systems that can coevolve with natural systems – evolve in a way that generates mutual benefits and greater overall expression of life and resilience. The field of regenerative development and design, which draws inspiration from the self-healing and self-organizing capacities of natural living systems, is increasingly seen as a source for achieving this end. This field is redefining the way that proponents of sustainability are thinking about and designing for the built environment, and even the role of architecture as a field.

Introduction

Chronology

Early Roots In the 1880s, Ebenezer Howard wrote *To-morrow: A Peaceful Path to Social Reform*. Reissued in 1902 as *Garden Cities of To-Morrow*, with an introductory essay by Lewis Mumford, the book was an early and influential expression of ecological thinking applied to human settlement. It sought to reconnect humans to nature, and featured use of natural rather than engineered processes to build the health of the system. His description of a utopian city in which man lives harmoniously together with the rest of nature stimulated the founding of the garden city movement and the establishment of several Garden Cities in Great Britain in the early twentieth century [11, 12].

In 1915, Patrick Geddes published his study of the urban growth patterns stimulated by the mass movement of people into cities [13]. Geddes, a biologist, saw cities as living organisms. He believed that addressing the problems of unsustainable growth required understanding a city's context – the surrounding landscape's natural features, processes, and resources – and called for a solid analytic method for developing that understanding. His conclusion would influence regional planning movements across Europe and the United States. Geddes applied the terms *Paleotechnic* and *Neotechnic* to distinguish the industrial era producing this destructive growth of human settlements from the era he predicted would follow its demise. These terms would be picked up by John Tillman Lyle, some 80 years later, to differentiate industrial era and regenerative technologies. Some trace the origins of ecological design to the work of Patrick Geddes [7, 11].

Development of the Ecosystem Concept and Ecological Perspective In 1935, Arthur Tansley introduced an entirely new concept to ecology in his work, "The Use and Abuse of Vegetational Concepts and Terms" [6]. He proposed the term *ecosystem* as a name for the interactive system of living things and their nonliving habitat and the application of systems science as a way to bring more scientific rigor to the study of nature's complexity and the effect of human activities on that complexity. Tansley and other organismic biologists of the period were the first to formulate a systems view of life. Seeking a more accurate depiction of how life ordered and organized itself within a particular landscape or geographic location, he posited that neither a living organism nor its physical environment could be thought of as separate entities: "we cannot separate them from their special environment, with which they form one physical system." Two of the most significant implications of this depiction of how life structures itself were the deconstruction of the human/nature dichotomy that had shaped Western design thinking and the establishment of the premise that all species are ecologically integrated with each other, as well as with the abiotic constituents of their biotope or habitat. For Tansley and other ecologists concerned about the increasing impact of humans on natural systems, the ecosystem offered a valuable framework for analyzing the effect of human activities on natural systems and resources. In later years, the concept was further defined or clarified to explicitly include a social complex (human social institutions and actions) and a built complex (structures and infrastructures) and became a framework for sustainable urban planning and development [14, 15].

In the 1950s and 1960s, Eugene and Howard Odum laid the foundation for the development of ecology into a modern science, based on the core concept of the ecosystem as the fundamental ordering structure of nature. They published the first textbook on ecology, *The Fundamentals of Ecology*, in 1953. Their work brought attention to the importance of understanding how the earth's ecological systems interact with one another. Howard Odum further developed a number of key theoretical concepts and methodologies including his "energy systems language," a set of symbols used to compose energy flow diagrams for any scale system [16]. His study of wetlands pioneered the now widespread approach of using wetlands as water quality improvement ecosystems and served as an important contribution to the beginnings of the field of ecological engineering [17].

New Foundations for Systems Theory and Systems Thinking In 1968, biologist and systems theoretician Ludwig von Bertalanffy published his *General System Theory: Foundations, Development, Applications*. General systems theory (GST) introduced the concept of open systems, emphasized the difference between physical and biological systems, and introduced evolutionary thinking – thinking focused on change, growth, and development [18]. GST opened the door to a new science of complexity. The recognition that complex systems cannot be understood through simple analysis led to the emergence of systems thinking as a major scientific field, a profound change from the analytic, reductionist mode that had dominated Western scientific thinking since the time of Descartes, Newton, Galisteo, and Bacon. GST also laid the basis for the development of living systems science, for Charles Krone's development of living systems thinking and their application to natural systems, as well as to human social systems. His work strongly influenced Howard Odum's ecosystem modeling and energetics, which, in turn, influenced John Tillman Lyle's work on regenerative design technologies.

In the 1960s–1970s – systems theorist and architect of organizational processes and structures, Charles Krone developed living systems thinking as a developmental technology for consciously improving systems thinking capacity. His work drew on and greatly extended GST and *Systematics*, a discipline developed by mathematician John Bennett that uses systemic frameworks to understand complex wholes within which people are participants rather than observers. The systemic frameworks and developmental processes Krone generated were applied and evolved within businesses. Their purpose was to create an understanding of businesses, communities, and nature as living systems and to build the consciousness required to create reciprocally beneficial relationships through better integration of industrial, community, and natural processes. His work served as a core foundation for Regenesis Collaborative Development Group as they developed and evolved regenerative development processes and technologies, starting in the 1990s [19, 20, 21]. Of particular importance for regenerative development was Krone's framework depicting four orders or levels of work that living systems of all scales need to carry out. Ranging over four levels from basic operations up through regenerative work, it allows practitioners to design for the integrated evolution of all levels of work in support of a regenerative change process.

Ecological Sustainability – Foundations of Regenerative Development and Design In 1969, Landscape architect Ian McHarg published *Design with Nature*, pioneering a technology for ecological land-use planning based on understanding natural systems [22]. His book became a foundational textbook for the ecological view of urban landscape design, and its basic concepts were later developed into the geographic information systems (GIS) – a critical tool for ecological development.

In 1978, Bill Mollison, an Australian ecologist, and one of his students David Holmgren coined the word permaculture from a contraction of *permanent agriculture* or *permanent culture*. They developed the field of permaculture as an ecological design system to promote design of human habitats and food production systems based on the relationships and processes found in natural ecological communities. Much of its inspiration was drawn from the relationships and adaptations of indigenous peoples to their ecosystems [9]. Like ecological practitioners such as Ian McHarg, Mollison and Holmgren espoused integration of human and natural environments, but they also developed design technologies and practices for increasingly self-sufficient human communities and food production systems. By creating "man-made ecosystems," permaculture demonstrated how to provide for a host of human needs while reducing dependence on environmentally destructive industrial practices. While earlier iterations of ecological design promoted integration of human and natural systems for more sustainable development, permaculture was the first ecological design system to introduce the concept of a regenerative effect as a new standard of ecological performance for the built environment – the generation of a surplus or overabundance of energy and resources that could be reinvested to evolve natural and human living systems as an integrated whole. In support of that goal, Mollison's *Permaculture: a designers' manual*, published in 1988, introduced a hierarchy of investment (regenerative, generative, and degenerative) as a framework for assessing the value of potential actions for building regenerative capacity in a system [9].

Also in the 1980s, Robert Rodale, son of organic agriculture pioneer J. I. Rodale, advanced the use of the word *regenerative* in relation to the use of land, calling for going "beyond sustainability, to renew and to regenerate our agricultural resources" [23]. Rodale used the term to describe the continuing organic renewal of the complex living system that he saw as the basis for healthy soil and, in turn, for healthy food and healthy people. He later applied the same principle of ongoing self-renewal to regenerative economic development [24]. While his work did not extend to the built environment, his principles influenced John Tillman Lyle's work and are foundational in the subsequent conceptualization and application of regenerative methodologies to all of the systems that support life.

In 1984, John Tillman Lyle published *Design of Human Ecosystems* [25] in which he argued that "designers must understand ecological order operating at a variety of scales and link this understanding to human values if we are to create durable, responsible, beneficial designs." He defined human ecosystems as "places in which human beings and nature might be brought together again" for mutual benefit and posited conscious ecosystemic design as essential to a sustainable future. The book introduced several key concepts that laid the basis for his subsequent work on regenerative design (below). These included (1) "shaping ecosystems, just like shaping buildings," requires a set of organizing principles drawn from "strong concepts of an underlying order that holds the diverse pieces and all their hidden relations together." (2) In ecosystem design, "these underlying concepts of order are drawn from ecology" and principles for ecosystem design draw from the "need to comprehend and envision the ecosystem the designer is seeking to shape as a dynamic (living) whole," and (3) ecological concepts are "more or less analogous to the laws of mechanics in architecture in that they provide us with organizing principles for shaping ecosystems much as architects shape buildings."

Ecological Design Systems Proliferate The 1990s was a period of intense creative ferment for ecological design thinking. A number of foundational books were published laying out both the practical and theoretical bases of design for ecological sustainability, including *Ecological literacy: education and the transition to a postmodern world* by David Orr (1992), *From Eco-Cities to Living Machines: Principles of Ecological Design* by Nancy Jack Todd & John Todd (1993), *The web of life: A new scientific understanding of living systems* by Fritjoff Capra (1995), *Ecological Design* by Sim van der Ryn and Stuart Cowan (1996), and *The ecology of place: Planning for environment, economy, and community* by Timothy Beatley (1997).

In 1992, educator David Orr and physicist Fritjof Capra coined the term ecological literacy (also referred to as *ecoliteracy*) to describe the ability to understand the natural systems that make life on earth possible, including understanding the principles of organization of ecological communities (i.e., ecosystems) and using those principles for creating sustainable human communities [4].

Also in the 1990s, new ecological and living system-based metric systems were introduced, including the revision of architect Malcolm Wells' Wilderness-Based Check-list for Design and Construction 1999 by the Society of Building Science Educators (SBSE) to address site and building issues. Their work acknowledged John Tillman Lyle's idea that sustainable design is merely breaking even, while regenerative design renews earth resources [20]. On a larger scale, Pliny Fisk's work on EcoBalance land-use planning and design method employed the principle of life cycles as a framework for sustaining basic life supporting systems, balancing human needs with their ability to manage the environment using technologies for augmenting natural processes [26].

Emergence of Regenerative Development and Design as Distinct Disciplines In 1996, John Tillman Lyle published *Regenerative Design for Sustainable Development*, the first comprehensive articulation of, and handbook for, regenerative design [11]. Written as a practical guide to the theory and design of regenerative systems, it laid out the framework, principles, and strategies for a design technology aimed at reversing the environmental damage caused by what Lyle called industrial land-use practices. The book reflected the continuing evolution of the thinking he had been pursuing as a landscape architect, architect and educator. He established the Center for Regenerative Design at California State Polytechnic University, Pomona to test, demonstrate, and further evolve this technology.

Deeply concerned about conventional industrial development's resource depletion and environmental degradation – consequences embedded in "the design of our twentieth century landscape," Lyle believed that at the core of growing environmental crises lay the simplification of living systems caused by "paleo" design and technologies (a term he adopted from Patrick Geddes to depict their relative crudity). "Where nature evolved an ever-varying, endlessly complex network of unique places adapted to local conditions," he wrote, ". . .humans have designed readily manageable uniformity." This creates relatively simple patterns and forms designed to be easily replicable anywhere. Most important, in his view, was the replacement of nature's continual cycling and recycling of materials and energy – processes "core to the earth's operating system" – with one-way linear flows from source to sink. "Eventually a one-way system destroys the landscapes on which it depends," Lyle observed. "The clock is always running and the flows always approaching the time when they can flow no more. In its very essence, this is a degenerative system, devouring its own sources of sustenance." The degenerative patterns caused by these linear, one-way flows he believed, demanded a fundamentally different approach that he named regenerative design. Accordingly, Lyle defined regenerative design as the replacement of linear systems of throughput flows with "cyclical flows at sources, consumption centers, and sinks." The resulting systems provide for "continuous replacement, through (their) own functional processes, of the energy and materials used in their operation" [11].

Lyle died just 4 years after publication of *Regenerative Design for Sustainable Development*. While he called redesign of the degenerative systems created by industrial linear flows as the "first order of work," it is clear from the larger body of his work and other writings [19] that he saw regenerative design as encompassing far more than this basic operational goal, as fundamental as it was. While much attention has been given to his models and techniques for designing self-renewing resource and energy flows, Lyle always saw the heart of his work, and the work of regenerative design, as the conscious design of whole ecosystems. His concern with the importance of developing a different nature of thinking as the basis for regenerative design, which was addressed in introductory chapters of the book, was also left without further development. Unfortunately, the narrow definition of the term regenerative as simply "self-renewing" came to define the focus of regenerative design for many architects and landscape architects.

In 1995, the authors of this chapter and the Regenesis Collaborative Development Group began developing the theoretical and technological foundation for regenerative development – enabling human communities to coevolve with the natural living systems they inhabit while continuously regenerating environments and cultures. Regenesis founders had practiced biocentric design, inspired by natural processes, in a variety of arenas for a number of years and knew the power of this approach. They agreed that development projects needed to be sources of ecological health, even "engines of positive or evolutionary change for the systems into which they are built" [27]. While agreeing with the ends of ecological sustainability, they felt that the primary cause driving unsustainable patterns was not being addressed by ecological design systems. They saw environmental problems as symptoms of a fractured relationship between people and the living web of nature, and argued that the core issue was cultural and psychological, rather than technological. Like Lyle, they believed that addressing this issue required a fundamental transformation in how humans saw their relationship and role with regard to the planet – moving from the current view of standing apart from and using (or protecting) nature to seeing a "coevolutionary whole, where humans exist in symbiotic relationship with the living lands they inhabit" [28].

For regenerative design to take hold and be successfully applied, they reasoned, a radical shift in thinking and understanding would be required among design professionals, stakeholders, and all the human inhabitants of a place. They proposed the term regenerative development for the more comprehensive work of creating the conditions and building the capacities required for achieving this shift, with the aim of making development a source of harmonious integration with nature [27, 29].

Regenerative Development and Design– Redefining Sustainability

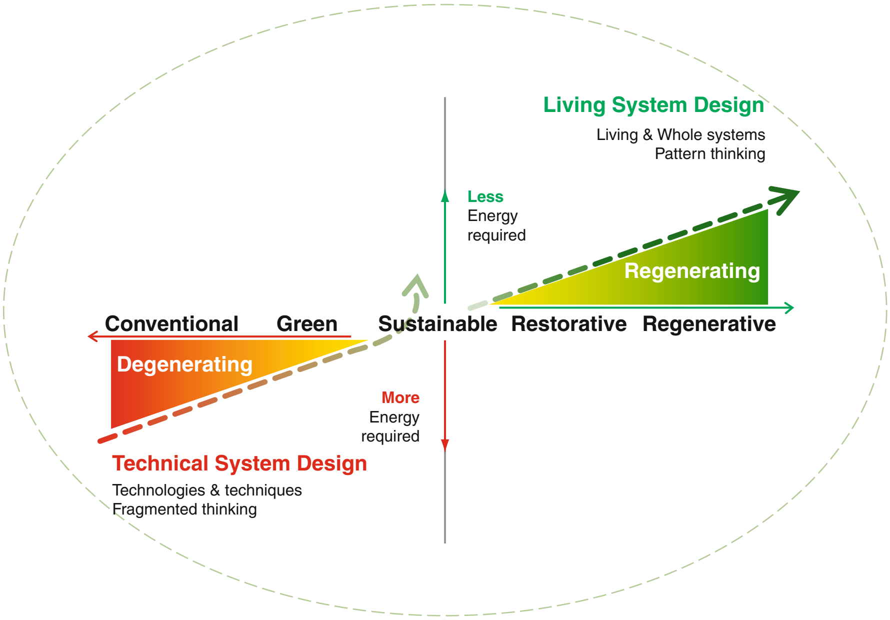
Introduction

Sustainable development and design has been described as falling broadly into two streams – one primarily technical and engineering based (technological sustainability) and the other based in ecology and living systems principles (ecological sustainability) [4, 7]. Green or high-performance building, sometimes called eco-efficient design, emerged out of the first stream, and regenerative development and design out of the second. Green building, like the conventional building field before it, defined the built environment as "all the structures people have built when considered as separate from the natural environment" (MacMillan Dictionary). It defined a sustainable built environment as one that is resource efficient and has minimal or neutral environmental impact. While that definition is evolving, the primary aim of green building continues to be increasing the efficiency of energy, water, and material use while reducing local and global impacts on the natural environment.

However, as Sarah Jenkin and Maibritt Pedersen Zari note in their 2009 research paper, "Rethinking the Built Environment," "The definition of a sustainable built environment is changing rapidly. While aiming for neutral or reduced environmental impacts in terms of energy, carbon, waste or water are worthwhile targets, it is becoming clear that the built environment must go beyond this. It must have net positive environmental benefits for the living world" [10].

The rising field of regenerative development and design, which emerged from the ecological stream, is not only leading the charge to redefine sustainability but also redefining what the built environment encompasses and what its role must be. Advocates of a regenerative approach to the built environment believe that a much more comprehensive, deeply integrated, and whole-systems approach is needed. They propose that eco-efficient design technologies and strategies be integrated within an ecologically based approach that reverses the degeneration of both the earth's natural systems and the human systems that inhabit them. The methodology of this approach focuses on the development of human settlements that partner with natural systems and processes to actively regenerate the health of their place as a whole and the spirit of the people who inhabit it (Fig. 1).

The philosophical and technical foundations for regenerative development and design as a distinctive field within ecological sustainability were laid in the 1990s, though they draw from scientific and technological advances reaching back into the early part of the last century (see "Chronology"). The practices emerging from that body of work are still evolving and expanding to cover an increasingly broad and sophisticated spectrum of sustainability concerns. Held together by a common philosophical core, they extend beyond the traditional aspects of design to address the



Regenerative Development and Design. Figure 1
Graphic contrast of Technical System Design and Living System Design. © Regenesis Group. Reprinted with permission

different nature of thinking and interactivity that is required to design and engage in a regenerative process.

While regenerative approaches are attracting growing interest among sustainability design practitioners, transitioning from green building to a regenerative practice has presented a number of challenges. The holistic and deeply integrated nature of the regenerative approach does not lend itself to a "menu approach" – picking one or two regenerative technologies without understanding the underlying principles that assure a regenerative outcome. Another challenge is reconciling the two radically different worldviews shaping technological and ecological sustainability within the way one's practice is carried out. Few architects and engineers are familiar with, let alone trained in an ecological paradigm. Yet as David Orr notes:.

► *Ecological problems are in many ways design problems: our cities, cars, houses, and technologies often do not fit in the biosphere. Ecological design requires the ability to comprehend patterns that connect, which means looking beyond the boxes we call disciplines to see things in their larger context. Ecological design is the careful meshing of human purposes with the larger patterns and flows of the natural world; it is the careful study of those patterns and flows to inform human purposes. Competence in ecological design requires spreading ecological intelligence—knowledge about how nature works.* [30]

Still, another challenge lies in the lack of a universally agreed upon definition for regenerative development and design and a tendency to blur or confuse regenerative approaches with the range of other design systems that emerged in pursuit of ecological sustainability in the 1990s.

This confusion around the distinctive nature of regenerative development and design has been due in part to being an emerging field lacking universal understanding of the meaning of regeneration, especially as it applies to design of the built environment. The distinction between regenerative development and regenerative design must also be further clarified if this field's potential contribution to sustainability is to be fully realized.

Overview: Ecological Sustainability and Regenerative Development and Design

Ecological Sustainability

Ecological sustainability has been defined as the "capacity of ecosystems to maintain their essential functions and processes, and retain their biodiversity in full measure over the long-term" (www.businessdictionary.com).

While accurate and straightforward, the seeming simplicity of this definition is deceptive. To understand and then deliver what is required to "maintain" and "retain" requires first understanding the nature of ecosystems and the nature of the ecological world in which they exist. That, in turn, requires understanding the ecological perspective – the use of ecological concepts from biology as a metaphor for understanding and designing environments.

All development of the built environment involves changing the landscape and, perforce, the natural systems embedded within it – modifying and adapting them for human purposes. The design of that change is ultimately based on the designer's understanding of the "nature of nature"– how nature works and, concomitantly, humans' relationship to it. That understanding, in turn, is shaped by the fundamental model or paradigm held by the larger culture – the metaphor used to depict how the world works. In the same way, technologies reflect a culture and how it understands nature [7, 11, 25].

The divergent definitions of ecological and technological sustainability can be attributed in large part to their being grounded in very different worldviews. Ecological sustainability as a field and the design systems within it emerged from the profound shift in worldview that occurred over the last century as a result of advances in both the physical and biological sciences. Fritjof Capra has described this as a shift from the mechanistic worldview of Descartes and Newton. In that paradigm, the dominant metaphor for understanding the world (and all organisms within it) was that of a machine composed of separate parts. In contrast, the ecological worldview sees the world as a self-organizing, continuously evolving, interdependent web of living systems, and the concept of ecosystem is the dominant metaphor for understanding its workings. The ecosystem concept, as it has been evolved and informed by living systems science, has been particularly influential in shaping ecological and regenerative understanding of the world and the role of humans within it, with profound implications for sustainability and development [5, 14, 15].

The industrial era metaphor of machine was particularly influential in shaping much of the built environment in the developed world and continues to play a significant role even today. By the first decade of the twenty-first century, however, Le Corbusier's image of the modern house as a "machine for living" was being challenged by the image of living buildings and communities as ecosystems.

As the ecosystem emerged as a new "governing concept of relationship between humanity and nature" [14], it confronted some of the most basic premises of the technologies, processes, and goals of the design field at the time, including the role of buildings, the definition of the built environment, the role of designers, and even the role of humans on the planet. As designers concerned about sustainability began to explore the implications of this new paradigm, it became clear that new ways of thinking and working, along with new forms of technology and new standards of ecological performance were required. The ecological sustainability stream, and within that stream, regenerative development and design, grew out of the work of the pioneering designers, educators, and scientists who took up the challenge of changing their design practice, themselves, and ultimately, their field in order to meet these requirements.

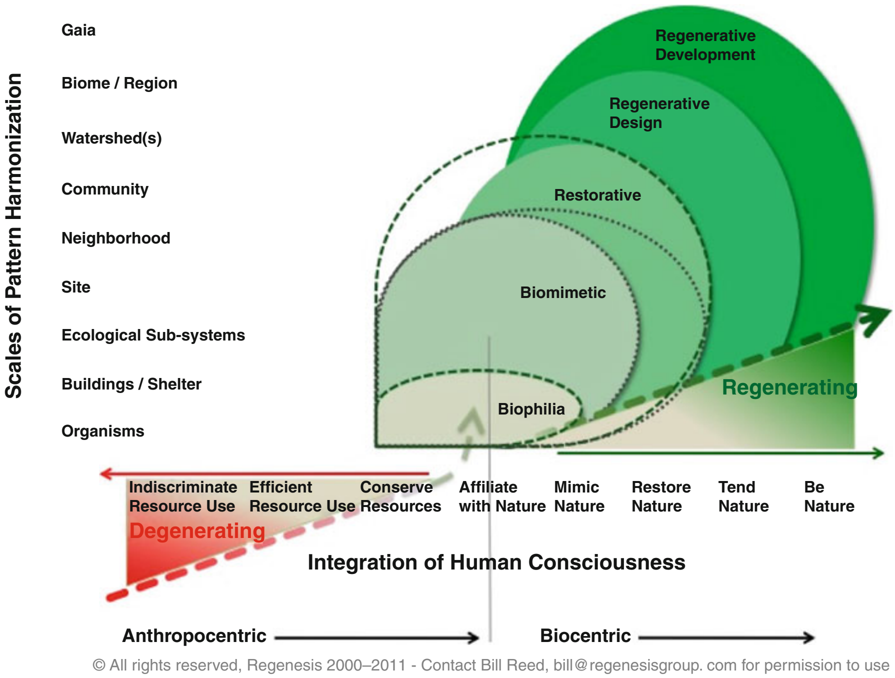
While many have written about different aspects of ecological sustainability during and since the 1990s, some of the most comprehensive articulations of the key premises that shaped the distinctive character of the broader field as a whole can be found in the writings of Sim Van der Ryn, Stuart Cowan, David Orr, and Fritjoff Capra [4, 5, 7, 30].

Regenerative Development and Design Overview

A number of ecological strategies for sustainability were developed during the 1980s and 1990s that were organized around the core set of philosophical, theoretical, and scientific concepts that underlie the ecological perspective of reality. All were aligned around a commitment to net positive goals for the built environment and to integrate human structures, processes, and infrastructures with natural living systems to that end. They differed in the systemic scope they encompassed, falling into four broad categories along a spectrum of comprehensiveness (Fig. 2).

1. Biomimetic – Cradle to Cradle and Biomimicry are design philosophies that fit into this category:

   Biomimetic approaches look to nature as inspiration. It is a *functional* approach that uses



Regenerative Development and Design. Figure 2
Levels of Ecological Strategies for Sustainability. © Regenesis Group. Reprinted with permission

nature – its forms and its processes – as a model for humans to follow – an anthropocentric perspective. Technical product design, buildings, manufacturing processes, agriculture, and human activity will function best and be more in harmony with ecological processes if nature is used as a model and guide. Nature's services and techniques are generally much more effective and certainly more sustainable than technical engineering approximations [31].

The principles guiding biomimetic thinking are essentially derived from an ecological understanding of how life works and provide a conceptual starting point to move into more whole and regenerative systems scope.

2. Biophilia – A general term, meaning "urge to affiliate with other forms of life" [32].

   As a design philosophy, biophilia is *relational* in its approach – it is somewhat passive in its engagement with life and is anthropocentric in its purpose. It acknowledges that humans will, if given a choice between nature and a human-made context, choose an environment or situation that utilizes, or is in contact with, living systems and their processes. Human health is positively influenced in relation to life and diminished if separated from living system connectivity. The design fields that use biophiliac approaches consciously use Literal Connections to natural features and elements; Facsimile Connections in terms of the use of nature imagery and materials; and Evocative Connections that use the qualities and attributes of nature in design such as sensory variability, prospect and refuge, serendipity, discovered complexity [33].

3. Restorative – Reestablishes the self-organizing and evolving capability of natural systems.

   This is an approach that acknowledges that humans have a role to play. It is more highly *integrated* than biomimetic approaches and more *active* than biophilic approaches – yet it generally is an episodic and finite engagement. This approach typically intervenes on an initial basis to reestablish the health of a subsystem of an ecosystem and community – such as wetlands, woods, riparian corridors, beach dune systems, social systems, and so on. It is a biocentric approach. When the intervening human role is finished, however – once the capacity of the system to self-organize is set in motion – the humans leave the engagement and are expected to [34].

4. Regenerative

   Acknowledges that humans are "nature," and there is greater hope of evolutionary potential in a state of intentional interrelationship. Humans have a positive role to play in nature. In order to create sustained ecological health, humans must evolve a conscious and *integral* interrelationship where humans and nature are in a mutually beneficial being and becoming relationship – one that is always aware of evolutionary potential. It is a fully conscious awareness that the health of an ecosystem is dependent on human health and human health is dependent on the health of the whole ecology. It is coevolutionary. This might be termed a process of *biobecoming* – the development of a whole system of interrelated living consciousness – a new mind. "Design inevitably instructs us about our relationships to nature and people that makes us more or less mindful and more or less ecologically competent. The ultimate object of design is not artifacts, buildings, or landscapes, but human minds" [4].

M. Kat Anderson supports this way of being in "Tending the Wild": *Wilderness is a negative label for land that has not been taken care of by humans for a long time...California Indians believe that when humans are gone from an area long enough, they lose the practical knowledge about correct interaction, and the plants and animals retreat spiritually from the earth or hide from humans. When intimate interaction ceases, the continuity of knowledge passed down through generations, is broken, and the land becomes "wilderness"* [35].

Together, regenerative development and design provide a framework for creating, applying, adapting, and integrating a blend of modern and ancient technologies to the design, management, and continuing evolution of sustainable built environments, accomplishing positive ecological and social results that include:

- Improving the health and vitality of human and natural communities – physical, psychological, economic, and ecological.
- Producing and reinvesting surplus resources and energy to build the capacity of the underlying relationships and support systems of a place needed for resilience and continuing evolution of those communities.
- Creating a field of caring, commitment, and deep connection to place that enables the changes required for the above to take place and to endure and evolve through time [10, 28].

The first comprehensive articulation of the theoretical and practical basis of regenerative approaches to the built environment emerged separately for regenerative development and regenerative design in the mid-1990s from two separate sources – the work of Regenesis Collaborative Development Group and John Tillman Lyle. Their respective bodies of work each reflected a convergence of disciplines in addition to architecture including, landscape ecology, geohydrology, landscape architecture, permaculture, regenerative agriculture, general systems theory and cybernetics, living systems theory and thinking, and developmental psychology.

Regenerative development and design, as articulated by Regenesis and Lyle, recognizes that "humans, human developments, social structures and cultural concerns are an inherent part of ecosystems" [10], making humans integral and particularly influential participants in the health and destiny of the earth's web of living systems. According to this view, the sustainability of the real estate development industry, which works directly on these webs, is largely determined by whether humans participate in them as partners or as exploiters [11, 25, 27, 28, 36, 37].

In his paper, "New Context, New Responsibilities: Building Capability" [38], Ray Cole articulated some of the key implications of a regenerative approach, including:

- Seeing the responsibility of design as "designing the 'capability' of the constructed world to support the positive coevolution of human and natural systems" versus designing "things" (buildings, infrastructure, etc.) and defining sustainable buildings as "buildings that can support sustainable patterns of living."
- Emphasizing the "role of building in positively supporting human and natural *processes*" versus "building as *product*."
- Positioning "building as central in creating higher levels of order and, as such, creating increased variety and complexity."

- Seeing the building as within and connected to a larger system – place, shifts "the current emphasis of greater energy self-reliance at the individual building level" to "opportunities for positive connections and creative synergies with adjacent buildings and surrounding natural systems."

A Note on the Distinction Between Regenerative Development and Regenerative Design For ecological sustainability to succeed, it requires a far broader and deeper scope of engagement than an individual building or even community design [39]. Yet the structure of the development and construction industry, for the most part, works to narrow the designers' role and scope, often as a result of decisions made before the design process even begins. Regenerative development was developed, in part, to address this concern. Regenerative approaches view development and design as two distinct yet synergistic processes, both of which play an essential role in ensuring that greater scope, neither of which is sufficient without the other.

The following dictionary definitions provide insight into the different roles of development and design:

Development: *O.Fr. desveloper, "an unfolding, bringing out the latent possibilities," from des- "undo" + veloper "wrap up"* a state in which things are improving; the act of improving by expanding or enlarging or refining; progression from a simpler or lower to a more advanced, mature, or complex form or stage; an unfolding; the discovering of something secret or withheld from the knowledge of others; disclosure.

Design: *L. designare "mark out, devise," from de- "out" + signare "to mark,"* an act of working out the form of something; to create or contrive for a particular purpose or effect.

Jenkin and Pedersen Zari, in their study, "Rethinking the Built Environment," write that "Regenerative development . . . investigates how humans can participate in ecosystems through development, to create optimum health for both human communities (physically, psychologically, socially, culturally and economically) and other living organisms and systems." They describe regenerative development as defining the desired outcome, and regenerative design as the means of achieving it. In contrast, John Tillman Lyle [25] defined design within the context of the built environment as giving form to physical processes [ref], and regenerative design as the replacement of linear systems of throughput flows with "cyclical flows at sources, consumption centers, and sinks." The resulting systems provide for "continuous replacement, through (their) own functional processes, of the energy and materials used in their operation" [11].

Regenerative development works at the intersection of understanding and intention, generating the patterned, whole-system understanding of a place, and developing the strategic, systemic thinking capacities and the stakeholder engagement required to ensure the design process achieves maximum systemic leverage and support. To that end, it integrates building, human, and natural development processes within the context of place. Regenerative development also creates an environment that greatly enhances the effect and effectiveness of restorative and biomimetic designs.

The roles of regenerative development, more specifically, are to:

1. Determine the *right* phenomena to work on, or to give form to, in order to inform and provide direction for regenerative design solutions that can realize the greatest systemic potential.
2. Build a field of commitment and caring in which stakeholders step forward as cocreators and ongoing stewards of those solutions.

Regenerative design then follows directly to offer a system of technologies and strategies based on an understanding of the inner working of ecosystems. Regenerative design solutions regenerate rather than deplete underlying life support systems and resources, are grown from the uniqueness of place, and work to integrate the flows and structures of the built and natural world "across multiple levels of scale, reflecting the influence of larger scales on smaller scales and smaller on larger" [40].

Regenerative Approaches to Sustainable Development and Design – Key Framework Premises and Practice Methodologies Overview

Key Premises

The following four premises are drawn from the work of Regenesis and Lyle. They offer key elements for framing regenerative approaches [11, 25, 27, 29, 37]. The four premises work as a system to integrate and align motivation and means, providing the framework within which methodologies and approaches from other ecological design systems can be integrated into a regenerative practice (Fig. 3). The first two define and shape motive and motivation in a regenerative project. The last two relate to how a project is carried out to ensure that ends and means stay congruent and that the process stays on course toward a regenerative result.

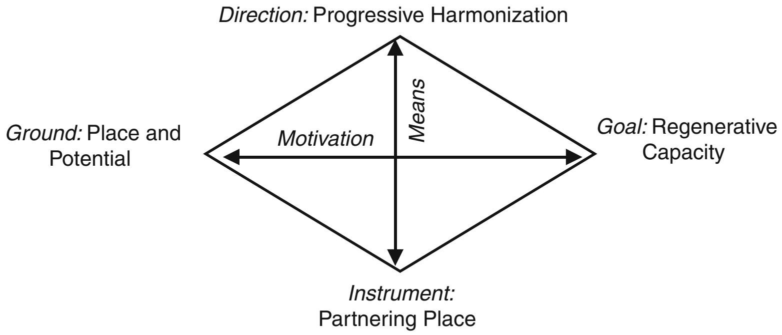
1. *Place and potential* – Understanding and conceptualizing right relationship to place. Starting with the richest possible understanding of the evolutionary dynamics of a place in order to identify the potential for realizing greater health and viability as a result of human presence in that place [41].
2. *Goals focus on regenerative capacity.* Regenerative project goals are defined by the capacity that must be developed and locally embedded to support ongoing coevolution of the built, cultural and natural environments, and the humans who utilize and tend to them toward higher (more complex, diverse, and generative) levels of order for all their constituent members as well as for the larger systems they are a part of and depend on [12, 37].
3. *Partnering with place.* Implementing a regenerative project requires taking on a new role, moving from a "builder of systems we control" to a gardener, working in partnership with a place and its processes [42].
4. *Progressive harmonization.* Regenerative approaches seek to catalyze a process of continually increasing the pattern harmony between human and natural systems and require indicators and metrics that can track dynamic, holistic, and evolving processes [43].

Place and Potential

- Potential: "the inherent capacity for growth, development or coming into being." (American Heritage Dictionary of the English Language)

William McDonough often describes design as an expression of human intention. Both that intention and the resultant design, however, are shaped by the potential the designer sees and seeks to realize for a particular project. Regenerative potential is defined as the ability to leverage human interventions to achieve greater systemic health through time for the place they occupy and depend on [27].

Many projects fail to achieve a regenerative effect because the potential they target is too limited – focused on an element or a problem without seeing its systemic connections. Others fail because they seek to realize potential defined by human ideals but fail to recognize and are thus unable to align with the essence of a place and the larger patterns of life that make it work. When a project is grounded in a rich patterned understanding of its place and a vision of its role and potential within that place guides its design, even small interventions can ripple out into large systemic transformations – what Curitiba's long-time mayor Jaime Lerner called "urban acupuncture" [44], with ecological as well as social and economic ramifications.


Direction: Progressive Harmonization
Ground: Place and Potential
Motivation
Means
Goal: Regenerative Capacity
Instrument: Partnering Place

Regenerative Development and Design. Figure 3
Framework depicting key premises and processes characterizing regenerative approaches. © Regenesis Group. Reprinted with permission

"Place" in regenerative development is alive, a living system or entity that is "…a unique constellation of patterns nested within patterns, interwoven with other patterns in families and guilds and social relationships, all endlessly changing, cycling, evolving and building to greater levels of complexity over time … an incredibly dynamic and complex being" [43]. A unique, multilayered dynamic network of natural and human ecosystems within a geographic region, this network forms a socioecological whole that is the result of complex interactions through time between and within its constituent ecosystems. The natural ecosystems include wildlife and vegetation, local climate, mineral and other deposits, soil, water geologic structures, etc.; human ecosystems include distinctive customs, expressions of values, economic activities, forms of association, ideas for education, traditions, physical artifacts such as buildings and constructed infrastructure, etc. [11, 28, 36, 37, 45, 46].

Regenerative Capacity: *Defining Goals for Realizing Regenerative Potential* The central element for regenerative development and design is the performance not of a single building, but rather of its living context – the unique socioecological system or "place" in which the building is just one of many interdependent and interactive elements and dynamics. Within that context, regenerative goals are set, and performance measured in terms of the intended contribution of the built environment to the regenerative capacity of that larger living context – (i.e., its capacity to realize and express more of its full potential as a source of increasingly healthy life for all its constituent members as well as for the larger systems it is a part of and depends on).

Characteristics of regenerative goals include:

- Place-sourced and place specific.
- Evolutionary, going beyond improving current systemic performance (what is often called restorative) to embedding into the system the capacity to continue to improve performance through time and through varying environmental conditions.
- Goes beyond functional performance goals. Recognizing "human aspiration and will as the ultimate sustaining source of our activities" [28], they address qualitative and spirit dimensions that shape the quality and degree of caring humans bring to their place and its capacity to continue to thrive.
- Focus on the processes physical structures enable as central.

Growing Capacity Versus Producing Things Regenerative projects set place and project specific goals that address all three aspects of regenerative built environments:

- Operational capacity.
- Organizational capacity.
- Aspirational capacity.

Operational capacity goals: Operational goals focus on systemic functional effectiveness in growing the potential of the underlying resource base – energy, materials, and support systems that enable the evolution of life in a place. Regenerative projects set goals for ensuring that the energies and nutrients flowing through it are used and invested optimally to grow the health of the system and all the life it supports.

Organizational capacity goals: Organizational capacity focuses on "who" a place is and addresses two dimensions – what is core to how this place works as a living system (what one can "mess" with and what one cannot) and what is the core qualitative character (its essence or distinctiveness, not data alone) or nature that humans can connect to at a heart level. Goals for this aspect deal with how to utilize the built environment and the design process to both illuminate and enhance the distinctive character of a place as something to be cherished. Historic codes and zones are often used to this end, but they tend to focus on surface appearance rather than essence, and over time, the code and its restrictions come to take the center stage, overshadowing the living core of the place they intended to protect [38].

Aspirational goals: Growing the systemic regenerative capacity of a place requires an integration of human aspirations with the distinctive ecosystems of that place and their drive to evolve their own health and generativity. This means harnessing inherent human creativity and aligning it with the creativity of nature and creating opportunities for people to experience themselves as able to make significant and meaningful contributions to their place [11, 47].

Partnering with Place – A New Role for Humans and Buildings In an ecological paradigm, sustainability requires a fundamental shift in how humans conceive of and carry out their role on the planet. In the words of Joshua Ramo, people must "change the role we imagine for ourselves from architects of a system we can control and manage to gardeners in a living, shifting ecosystem. For hundreds of years now we have lived in our minds as builders: constructing everything from nations to bridges . . . In a revolutionary age, with rapid change all around us, our architects' tools are deadly. It is time for us to put them down and follow (Nobel Laureate Friedrich von) Hayek's injunction to live and to think as gardeners."– gardeners who see themselves as partners in coevolution with the living system in which they work, cultivating "growth by providing the appropriate environment, in the manner a gardener does for his plants" [42, 48].

Successful regenerative development ultimately requires all the stakeholders in a place, not just the development/design team to move from the role of "builder" to "partner-gardener," with the first step a different nature of understanding that enables people to see the places they inhabit as alive.

A whole-systems assessment looks at a wide range of patterns covering multiple scales of systems and a number of different facets. The place intelligence it develops is a resource that can be mined to inform each stage of design to help ensure that the patterns generated by the project harmonize with the larger patterns of place. Another nature of understanding, however, is required to generate the experience of connection and caring that creates a relationship of partnership with a place. This understanding conveys "who" a place is as a living being in addition to how it functions. Every living system – whether a person, a tree, or a place – has an ongoing and distinctive core from which it organizes the complex arrays of relationships that produce its activities, its growth, and its evolution. Being able to grasp and share the distinctive core or *essence* of a place among and between the design team and local stakeholders provides an enduring basis for strong partnering relationships; in the same way, it builds strong human partnerships.

▸ A New Way of Thinking:

Learning how to apply a regenerative approach begins not with a change of techniques but rather with a change of mind—a new way of thinking about how we plan, design, construct, and operate our built environment [27].

Growing stakeholders and designing and constructing projects that can work as "place gardeners" requires bringing and developing whole-systems thinking that is able to and capable of comprehending and ordering and organizing the systemic complexity and dynamism of a living place and its multiple scales of nested systems, interactions of multidisciplinary teams over extended periods, and extensive local stakeholder participation [14, 15, 49]. This nature of systems thinking is characterized by:

- Being grounded in ecoliteracy and pattern literacy. Ecoliteracy applies an understanding the fundamental principles that govern how living systems work to specific situations and conditions. Pattern literacy involves being able to read, understand, and generate appropriate patterns that harmonize with and enable a place and its inhabitants to more fully realize what they can be [43].
- Requiring the practitioners to see what they are working on as a system of energies or life processes rather than as things – illuminating the constant reaching toward being more whole and being more alive inherent in living systems that is the fuel for regeneration [29, 50].
- Enabling a diversity of participants to grow their own systems thinking capacity in order to take on more challenging, value-adding roles [21, 50].

A new way of working: Regenerative development and design does not end with the delivery of the final drawings and approvals, or even with build out of a project. The responsibility of a regenerative designer includes putting in place, during the development and design process, what is required to ensure that the ongoing regenerative capacity of the project, and the people who inhabit and manage it, is sustained through time. Regenerative development employs *Developmental Design Processes*. They encompass integrative design (integrative, interdisciplinary beyond traditional building disciplines, open, and participatory) [51] and go beyond to embed self-managed learning processes into the work of conceptualizing, designing, constructing, managing, and evolving regenerative projects. They integrate the traditional focus of organizing for task accomplishment with the development of new thinking capacities required to design processes not things, make ecologically sound place-appropriate decisions, and create the being connection to and emotional resonance with place that generates the will required to follow through on those decisions.

Progressive Harmonization The "pole star" or overarching source of direction for regenerative projects derives from the ultimate effect every regenerative project seeks to achieve: an enduring and mutually beneficial relationship between the human and natural systems in a particular place. Pattern is the language of relationship, and regenerative development and design in a living system is a process of patterning human communities to align with the energetic patterns of a place in a way that both humans and the place coevolve. Christopher Alexander was speaking of pattern harmony when he wrote, "When you build a thing, you cannot merely build that thing in isolation, but must also repair the world around it, and within it, so that the large world at that one place becomes more coherent, and more whole; and the thing which you make takes its place in the web of nature, as you make it" [52]. While his initial work focused primarily on the pattern relationship between a building and the human community and life surrounding, his later work has increasingly encompassed all living systems. Wendell Berry, in his essay Solving for Pattern, speaks to creating pattern harmony between human communities and activities and the biosphere they take place in [53]. "A bad (design) solution is bad," Wendell Berry notes, "because it acts destructively upon the larger patterns in which it is contained. . . most likely, because it is formed in ignorance or disregard of them. A good solution is good because it is in harmony with those larger patterns . . . A bad solution acts within the larger pattern the way a disease or addiction acts within the body. A good solution acts within the larger pattern the way a healthy organ acts within the body" [53].

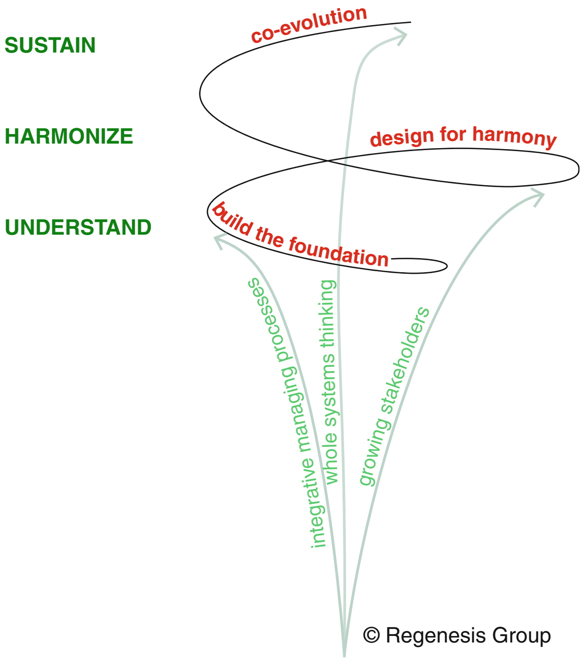
Pattern harmony, however, is not a stable state; a good solution today may become a bad one in a few years, so solving for pattern requires a progressive rather than one-time harmonization, a continuous repatterning. Theoretical biologist Stuart Kauffman called this mutually beneficial relationship "co-evolving mutualism" – co-evolving because "its ecosystems are always in the process of self-organization and reorganization, increasing in complexity, definition, and information content" [25, 54, 55].

Practice Methodologies

The following is an example of how these premises can translate into a regenerative practice. The methodologies were developed from over 15 years of fieldwork by Regenesis during which collaborative members explored, practiced, and evolved regenerative development. The diagram in Fig. 4 was developed as a depiction of the essential elements of this practice – three phases and three developmental processes that are considered key to creating and sustaining an evolutionary spiral, growing systemic capacity as it actualizes a project.

The Three Key Steps:

Understand the Relationship to Place: Integral assessment – a whole-systems (cultural, economic, geographic, climatic, and ecological) assessment of



Regenerative Development and Design. Figure 4 Regenerative practice methodology framework. © Regenesis Group. Reprinted with permission site and place as living systems lays the foundational understanding and thinking required to see how humans can enable the health and continuing evolution of the place and themselves as a part of it. A Story of Place® is codeveloped with the client and/or community. It uses the power of storytelling to articulate the essence of a place, how it fits in the world, and what the role of those who inhabit it can be as collaborators in its evolution.

Designing for Harmony with Place: Translates this understanding into design principles and systemic, integrated plans, designs and construction processes that optimize the presence of people in a landscape by harmonizing with the larger pattern of place. Buildings and infrastructure improve land and ecosystems, and the unique attributes of the land improve the built environment and those who inhabit it. Synergy with the land and ecosystems leverages the effectiveness of green design features and technologies and lowers costs while improving ecosystem health and productivity.

Co-Evolution: "*. . .sustainability means maintaining the dynamic potential for further evolution. Living systems survive by maintaining a condition of dynamic equilibrium with the environment through constant change and adaptation. In the game of evolution, equilibrium is death" Urban Sustainability Learning Group* [56]. This phase unfolds from the work of the previous two phases. If they have succeeded in creating a culture of coevolution in and around the project, and not just a physical product, its effect can be seen even before final build out. The role of designer becomes one of resource, providing processes and methods for sustaining the connection to place as a context that enables owners, managers and maintenance contractors, and community stakeholders to recognize and incorporate new social, economic, and ecological opportunities as their place evolves.

The Three Key Determining Factors: Success in the above three steps is determined by how one thinks, how one identifies harmonies and harmonizes the human role, and how one engages stakeholders *throughout* the planning and development process. Specifically, whether one:

- *Applies whole-systems thinking* to the design, planning, and decision making processes.
- *Manages integration and harmonization* across disciplines, between phases and team members and local stakeholders.
- *Grows stakeholders* understanding and appreciation of the place and the new potential offered and their capacity to be increasingly effective partners with the whole system of evolving life.

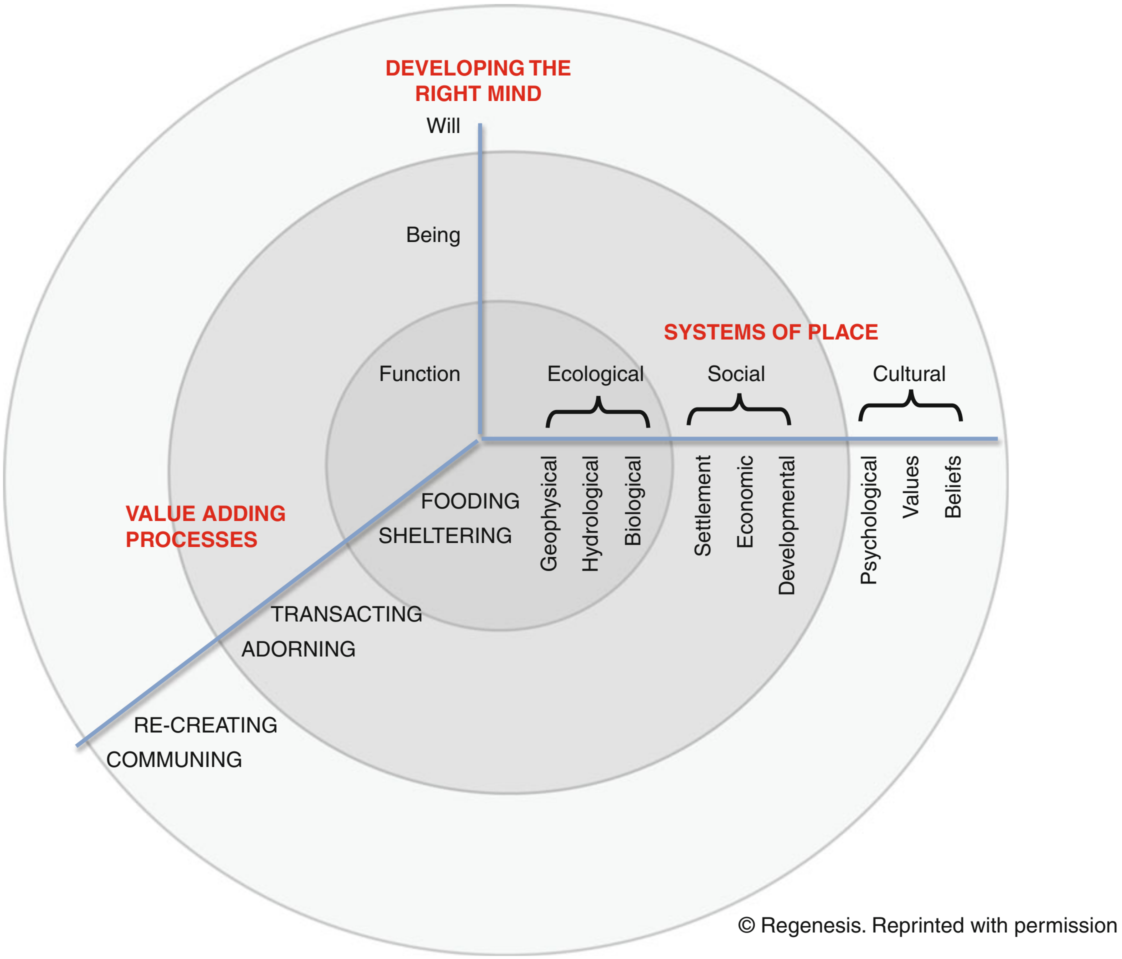
The following are examples of the thinking and practice frameworks and methodologies applied within the three phases and processes, some developed by Regenesis, some drawn from other ecological design systems.

Understand the Relationship to Place Principles from permaculture and Biomimicry are helpful in developing specific land use, building, and infrastructure design strategies.

Permaculture: *As* a design system rooted in the ability to discern the patterns that are structuring both natural and human systems and to generate new patterns that weave the human and natural together into a dynamic whole, permaculture assessment methodologies provide a source for developing holistic site assessments. *Pattern as Process*, an article by Regenesis principal Tim Murphy and Vickie Marvick, provides a detailed description of their method for understanding and interpreting the patterning of a site and its place [43].

Assessment Scope Framework for Pattern Understanding (Fig. 5): The challenge in any assessment process is to ensure that the scope being assessed is whole enough to encompass the interweaving of human and natural systems, dynamics and flows that shaped the distinct character of a place. Regenesis developed the following framework as a means of illuminating the core patterns structuring a place as the basis for "mapping" their dynamic and evolving interrelationships. These include:

- The ecological, social, and cultural systems creating and managing the conditions that shape how life expresses itself in a place.
- The value-adding processes life engages in within the context of those conditions and how they influence and are influenced by them.



© Regenesis. Reprinted with permission

Regenerative Development and Design. Figure 5
Integral Assessment Scope framework. Used as a means of illuminating the core patterns structuring a place as the basis for "mapping" their dynamic and evolving interrelationships. © Regenesis Group. Reprinted with permission

- The developmental implications and opportunities for how individuals (people and buildings) can enable the health and continuing evolution of the place and themselves through how they function, the qualitative state of being they seek and enable, and what they value and express will toward (Adapted from a framework developed by Charles Krone as part of his thinking technology [8]).

Essence Understanding: The essence understanding that conveys "who" a place is as a living being emerges from the whole-systems assessment. Questions used to reveal the essence include: What is at the core of a system, around which it is organized? What is the web or larger context of reciprocal relationships within which it is embedded since all systems are comprised of smaller systems and part of larger systems? And what is the potential inherent in a living system that it is attempting to live out since this is the fuel for regeneration – this constant reaching toward being more whole, being more alive?

► Simple example of patterns and the essence of a system: [51]

Mahogany Ridge, Idaho, USA: *A reductionist approach or an approach that abstracts life into a checklist might state that nothing should be built on existing farmland. This might be a good principle if the* *agriculture system was truly symbiotic with nature. In this case, farming had nearly destroyed three distinct ecological systems. An integral assessment looked for possible patterns of life that allowed for high levels of relationship between species and ecological niches.*

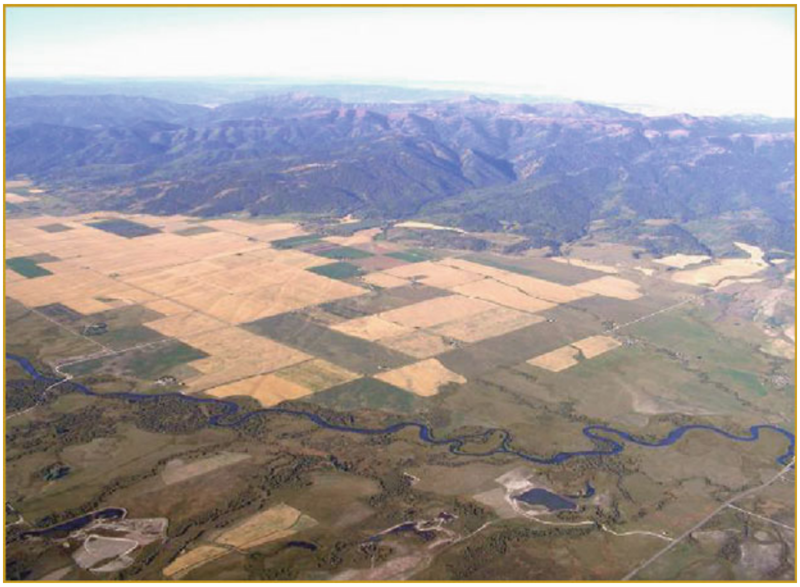
The aerial photo in Fig. 6 depicts approximately 3,500 acres of current farmland along the eastern edge of the Big Hole Mountains (just west of the Grand Tetons) that was being considered for development. Originally, these mountain watercourses and alluvial fan supported beaver, otter, native cutthroat trout, salmon, turkeys, grouse, and mega-fauna, such as deer, elk, moose, and bears. These animals were all responsible for carrying nutrients back upstream into the mountains to feed the forest and diversify the terrestrial and riparian ecosystem. Pioneers of European descent arrived in this place 100 years ago and used row-crop agriculture techniques to farm on this alluvial fan. As a result, ninety percent of the water from the Big Hole Mountains (in picture) was being used for agricultural purposes (spray irrigation), the salmon were no longer breeding in the river, the Yellow Tail cutthroat trout were in species decline, the river was polluted from overloads of nitrogen, and the upstream forests were in decline.

The area farmers were going out of business or bankrupt due to the short growing season. The farms, in the past, had been used to support local needs. Twenty to forty acre-per-home zoning is planned as the alternative to large farms.

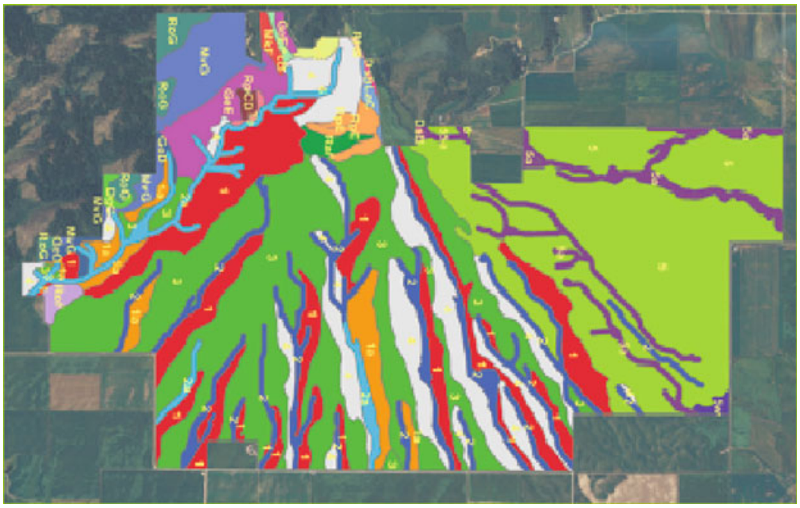
Looked at closely, this photo in Fig. 6 reveals that farming was superimposed on top of this alluvial fan between the stream in the mountain valley (top center of the photograph) and the river. The soils mapping indicated in Fig. 7 reveals the pattern more clearly.

Before farming took place here, these radiating streams and drainage ways served as additional corridors of cover for wildlife moving back and forth between the mountains and the river. When farmers settled the land, they diverted this perennial stream along the highest possible course (in elevation) to irrigate fields that were gridded over a highly productive and robust prairie ecosystem. This action severely simplified and destabilized the ecosystem that once was there. The farming pattern did not preserve the integrity of the pattern that contained it; rather, this larger healthy pattern was obliterated. The ecological function of this alluvial fan, and one of the *core patterns* of the ecosystem in this place, is that of a *"living bridge"* between the mountains to the west and the Teton River.

The pattern of a living nutrient bridge between the mountains and the valley that had been revealed in the assessment indicated that a higher level of potential



Regenerative Development and Design. Figure 6
Aerial photo of Mahogany Ridge Resort Community site



Regenerative Development and Design. Figure 7
Soil map of Mahogany Ridge site showing alluvial fan patterns

health can be reestablished in this mountain, alluvial fan, and River system. The development of homes in tight clusters could be used to pay for the restoration of the stream and habitat corridors that originally connected the Teton River and the mountains and provide wildlife corridors as well as many ecosystem services for community residents. To support the reestablishment of wildlife corridors, no fences would be allowed, native grasses would be planted (minimal turf grass), no off-leash dogs to disrupt nesting, and territory establishment by new wildlife.

By integrating the community into the development and management of these systems, they could produce food (through diversified agriculture and wild harvesting), timber, and other products, as well as the development of a diversified economy while insuring the provision of ecosystem services for their community. The human involvement in these patterns and processes is key to the ongoing regeneration and development of potential of the site.

Designing for Harmony with Place *Biomimicry*: The Biomimicry Guild's Life's Principles and their Genius of Place program provide guidance and models for establishing locally attuned strategies for building design through looking at how local species live out universal ecological principles within the conditions of a particular site and its surroundings, and how they have adapted to thrive within it (web link: www.biomimicryguild.com/).

Permaculture principles, which draw both on an understanding of ecology and of how indigenous people engaged with their place provide a lens for developing design strategies for responding to site conditions and opportunities in a way that is mutually beneficial (http://permacultureprinciples.com/; www.tagari.com/).

Two other useful frameworks that can be utilized as a part of the designing for harmony phase include:

A Regeneration-Based Checklist for Design and Construction (SBSE).

Malcolm Wells' Environmental Checklist.

www.sbse.org/resources/docs/wells_checklist_explanation.pdf

Essential Living Processes Framework: This framework was developed by Regenesis for setting overarching project aims to guide the design and construction process. It is based on the six critical processes that enable living systems to support the evolution of life. They include the ability to provide the material structuring that forms the basis for life processes – nourishment, shelter (habitat), and the generation and exchange of resources for growing and evolving more life. Because humans cannot be separated out from any living system, the factors go beyond the material factors – the outer landscape of a place. They also include the "inner landscape" that sources one's spirit and will and drives one to cherish and protect the places one inhabits. They include the ability of a living system to create a sense of identity and foster belonging through its culture, to support meaningful and contributory lives, and to invoke the spirit and inspiration that sustains caring. The framework enables setting aims and goals (and later developing indicators and measuring systems) for how the processes generated by the project support ecological, economic, and social health in each of the six areas. The interrelationship of these processes and how they cross ecological, societal, and economic arenas is graphically represented in Fig. 8:

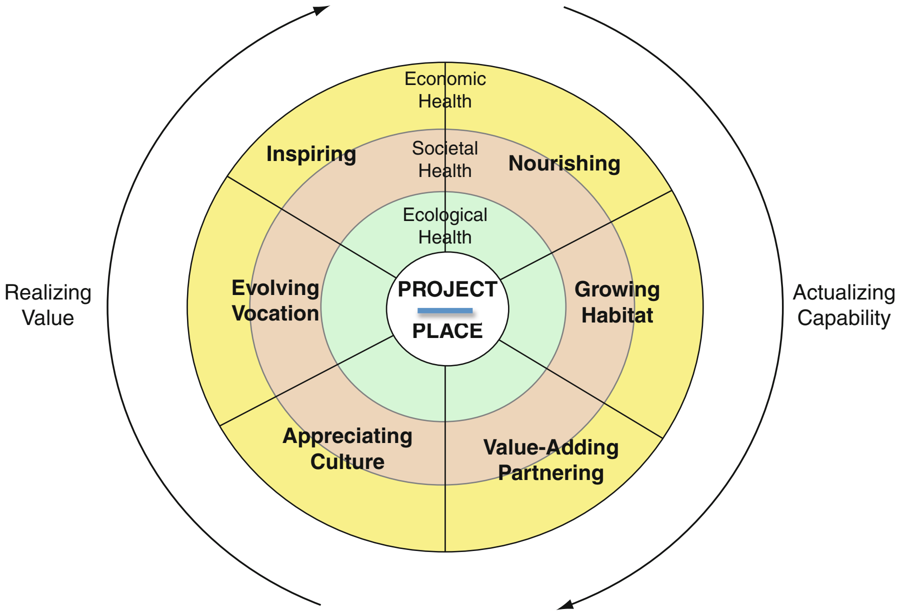
An example of ecological aims for nourishing:
Capacity of soil, water, and air to nourish life –
Aims:

- Invest water in increasingly higher order life processes through storing and cycling (vs. hoarding) so water becomes the driver for improving soil and air quality.
- Products and processes used in construction and operations are investments in growing the capacity of soil, water, and air to store, transform, and transport nutrients for optimal accessibility and utilization.

Future Directions

While regenerative development and design still occupies a relatively small niche in the larger world of sustainability efforts, interest in regenerative approaches to the built environment is on the rise. Beyond the USA itself, this growing interest has been particularly marked in Australia and New Zealand, including a government-commissioned research report that recommended the latter adopt regenerative development as a national policy [10].

A number of interrelated factors, working as a system, are creating a favorable climate that is likely to continue to feed such interest, among them: more practitioners encountering the limits of green building to address the global crises, shifting market dynamics and public awareness, the growing influence of the ecological perspective and the ecosystem concept, the movement toward integrative design with its reliance on interdisciplinary teams, and the growing recognition of the need for community



Regenerative Development and Design. Figure 8
Framework showing interrelationship of the 6 essential living processes and how they cross ecological, societal, and economic arenas. Used to set holistic, integrative goals and indicators. © Regenesis Group. Reprinted with permission

engagement and participation to support the behavior changes required for enduring sustainability.

In the 1990s, the most discussed issue for aspiring green designers was how to convince clients to incorporate sustainability features. By 2010, the discussions increasingly were about how to meet clients' demands for making their project "the greenest" of their kind. Over the same period, appreciation and understanding of ecological sustainability and the ecosystem perspective as it applies to human settlements and institutions has been significantly reshaping thinking in such fields as public health, education, economic and community development, and urban planning, as well as design of the built environment. Its core concepts, especially the concept of seeing communities as ecosystems in which nature and culture, human and natural designed features are interwoven and interdependent, are driving a move toward increasingly systemic and comprehensive goals. These comprehensive goals are, in turn, defining new standards of sustainability. Projects seeking to be "the greenest" now include social, economic, educational, and esthetic goals as well as goals around energy efficiency and pollution. More comprehensive goals affecting multiple fields are necessarily stimulating more integrative and interdisciplinary approaches. They are also adding the need to build community support and stewardship to the list of essential design issues. The ecological and ecosystem perspectives are providing a common "language" or set of frameworks across those fields that is facilitating integrative and participatory approaches across disciplines and between design teams and the public and in the process, further reinforcing an ecological worldview.

One effect of this system of factors has been the extension of explicitly regenerative approaches across a wider spectrum of fields and the integration of these fields with regenerative development of the built environment as part of regenerative community development. Regenerative development had already begun to shift the old, building-centric definition of the built environment to include the relationships between and among buildings, infrastructure, and natural systems, as well as the culture, economy, and politics of communities. Its concept of place-sourced design is providing a means of engaging the will of a community around aligning human and natural communities around shared purposes. Given its holistic and integrative character, it could be anticipated that these more comprehensive applications will be a continuing trend.

Regenerative Development makes possible a new and critically needed role for developers and developments, the full potential of which is still unfolding. Development projects already exist that, by the way they are built and occupy land, serve as instruments for reversing ecological damage, and as economic forces for constructing sustainable livelihoods. Still other projects offer glimpses of how, through weaving the many stories of Place into a mutually appreciating whole, a Regenerative Development becomes a harmonizing force within communities and among different stakeholders, inspiring new standards of appropriate relationship to Place. It is possible to envision how, by introducing larger systemic vision and potential, development can become a catalyst for the creation of self-evolving bioregional infrastructures and cultures of regeneration.

While this new role is beginning to emerge in small scales and at scattered locations, it is largely unrecognized as being part of a larger evolution. What is needed now is to bring consciousness and intention to its emergence as the new pattern shaping the field of development.

Bibliography

Primary Literature

1. Benyus J (1997) Biomimicry. Harper Collins, New York
2. McDonough W, Braungart M (2003) Cradle-to-cradle design and the principles of green design. www.mcdonough.com/writings/c2c_design.htm. Accessed 18 Oct 2011
3. McDonough W, Braungart M (2002) Cradle to cradle: remaking the way we make things. Northpoint, New York
4. Orr D (1992) Ecological literacy: education and the transition to a post-modern world. State University of New York Press, Albany
5. Capra F (1996) The web of life: a new scientific understanding of living systems. Anchor Books, New York
6. Tansley AG (1935) The use and abuse of vegetational concepts and terms. Ecology 16:284–307. doi:10.2307/1930070
7. Van der Ryn S, Cowan S (1996) Ecological design. Island Press, Washington, DC
8. Krone C (2001) West coast resource development: session notes. Unpublished transcription of dialogue by members of the Institute for Developmental Processes, Carmel
9. Mollison B (1988) Permaculture: a designers' manual. Tagari, Australia

10. Jenkin S, Zari MP (2009) Rethinking our built environments: towards a sustainable future. Ministry for the Environment, Manatu Mo Te Taiao, Wellington
11. Lyle JT (1994) Regenerative design for sustainable development. Wiley, Hoboken
12. Howard SE (2011) Encyclopædia Britannica. http://www.britannica.com/EBchecked/topic/273428/Sir-Ebenezer-Howard. Accessed 18 Oct 2011
13. Geddes P (1915) Cities in evolution. Williams & Norgate, London
14. Marcotullio PJ, Boyle G (eds) (2003) Defining an ecosystem approach to urban management and policy development. United Nations University Institute of Advanced Studies (UNU/IAS), Yokohama
15. Pickett STA, Grove JM (2009) Urban ecosystems: what would Tansley do? Urban Ecosyst 12:1–8. Published online: Springer Science + Business Media, LLC
16. Odum HT (1971) Environment society and power. Wiley Interscience, New York
17. Mitsch WJ, Jørgensen SE (1989) Ecological engineering: an introduction to ecotechnology. Wiley, New York
18. Von Bertalanffy L (1968) General system theory: foundations, development, applications. George Braziller, New York
19. Lyle JT (1993) Urban ecosystems. In context 35, p 43
20. SBSE/Wells (1999) www.sbse.org/resources/docs/wells_checklist_explanation.pdf. Accessed 18 Oct 2011
21. Sanford C (2011) The responsible corporation: reimagining sustainability and sustainability and success. Josey-Bass, San Francisco
22. McHarg IL (1969) Design with nature. Doubleday, Garden City
23. Rodale R (1988) Whole Earth Review, interview. http://findarticles.com/p/articles/mi_m1510/is_n61//ai_6896856/. Accessed 18 Oct 2011
24. Medard G, Pahl E, Shegda R, Rodale R (1985) Regenerating America: meeting the challenge of building local economies. Rodale, Emmaus
25. Lyle JT (1984) Designing human ecosystems. Wiley, Hoboken
26. Pliny F http://cmpbs.org. Accessed 18 Oct 2011
27. Haggard B, Reed B, Mang P (2006) Regenerative development. Revitalization, Mar/Apr 2006
28. Mang N (2009) Toward a regenerative psychology of urban planning, Saybrook Graduate School and Research Center, San Francisco. http://gradworks.umi.com/33/68/3368975.html. http://powersofplace.com/papers.htm. Accessed 18 Oct 2011
29. Haggard B (2002) Green to the power of three. Environ Des Constr, 24–31 Mar/Apr 2002
30. Orr D (1994) Earth in mind. Island Press, Washington, DC
31. Zari MP, Storey J (2007) An ecosystem based biomimetic theory for a regenerative built environment. In: Lisbon Sustainable Building Conference, Lisbon
32. Wilson EO (1984) Biophilia. Harvard University Press, Cambridge
33. Heerwagen J (2007) Biophilia and design. Handout for Portland lectures, Aug 2007
34. Kellert S (2004) Beyond LEED: From low environmental impact to restorative environmental design. Keynote address, greening rooftops for sustainable communities conference. Sponsored by Green Roofs for Healthy Cities, Toronto, and City of Portland, Portland, 4 June 2004
35. Anderson MK (2005) Tending the wild: native American knowledge and the management of California's natural resources. University of California Press, Berkeley
36. Cole RJ, Charest S, Schroeder S (2006) Beyond green: drawing on nature (for the Royal Architectural Institute of Canada's "Beyond green: adaptive, restorative and regenerative design" course – SDCB 305). The University of British Columbia
37. Reed B (2007) A livings systems approach to design. AIA National Convention May – Theme Keynote address
38. Cole R (2010) New context, new responsibilities: building capability. https://members.weforum.org/. . ./Ray%20Cole%20-%20Building%20Capability.pdf; http://www.google.com/url?sa=t&source=web&cd=1&ved=0CCIQxQEwAA&url=http%3A%2F%2Fdocs.google.com%2Fviewer%3Fa%3Dv%26q%3Dcache%3AaiaZ4YMGb0IJ%3Ahttps%3A%2F%2Fmembers.weforum.org%2Fpdf%2Fip%2Fec%2Fbreakthroughideas%2FRay%252520Cole%252520-%252520Building%252520Capability.pdf%2Bray%2Bcole%2BNew%2BContext%2C%2BNew%2BResponsibilities%3A%2BBuilding%2BCapability%26hl%3Den%26gl%3Dus%26pid%3Dbl%26srcid%3DADGEEShGS-VECswRMKKDehwApC-PqrSXfEB7LVbS2FGYn7eF7SuydqGa-76EwFbOdRXXa05jSIcKwPy13JevQLvd_Z3wD4lhOQVCZTjzJ63T-D9tcJXvvlDYO0OocQhRaJu3rlgk09fT0%26sig%3DAHIEtbRfn0IrDnH7Bqa2piU7yKiuWbpfzw&ei=4OadTtTNDPPaiQLjwZ3MCQ&usg=AFQjCNGXqaLa9z0t0FrfdQxeNvdHTia2ng&sig2=6x71Tx9WJSSjx5IjeDikLw. Accessed 18 Oct 2011
39. Williams D (2007) Sustainable design: ecology, architecture, and planning. Wiley, Hoboken
40. Bailey RG (2002) Ecoregion-based design for sustainability. Springer, New York
41. Mang N (2006) The rediscovery of place and our human role within it, Saybrook Graduate School and Research Center, San Francisco. http://powersofplace.com/papers.htm. Accessed 18 Oct 2011
42. Ramo J (2009) Age of the unthinkable: why the new world disorder constantly surprises us and what we can do about it. Little Brown and Company, New York
43. Marvick V, Murphy T (1998) Patterning as process. Permaculture Activist, Jul 1998
44. Lerner J (2005) Acupuntura Urbana. Institute for Advanced Architecture of Catalonia
45. Gabel M (2009) Regenerative development: going beyond sustainability. Design Science Lab, New York
46. Mang P (2001) Regenerative design: sustainable design's coming revolution. Design Intelligence. http://www.di.net/articles/archive/2043/. Accessed 18 Oct 2011
47. Orr D (2001) Architecture, ecological design, and human ecology. In: Proceedings of the 89th ACSA annual meeting. ACSA, Washington, DC, pp 23–32

48. von Hayek F (1974) The Pretence of Knowledge. Nobel Prize acceptance speech
49. Pickett STA, Cadenasso ML (2002) Ecosystem as a multidimensional concept: meaning, model and metaphor. Ecosystems 5:1–10. doi:10.1007/s10021-001-0051-y
50. Sanford C (2006) Building intelligence: a living systems view. Springhill, Battleground
51. 7 Group, Reed B (2009) The integrative design guide to green building: redefining the practice of sustainability. Wiley, Hoboken
52. Alexander C (1997) A pattern language: towns, buildings construction, Center for Environmental Structure Series. Oxford University Press, New York
53. Berry W (1981) Solving for pattern in gift of good land. Counterpoint, Berkeley
54. Kauffman S (2008) Reinventing the sacred: a new view of science, reason, and religion. Basic Books, New York
55. Prigogine I (1997) End of certainty. Free Press, New York
56. Urban Sustainability Learning Group (1996) Staying in the game: exploring options for urban sustainability, Tides Foundation Project

Books and Reviews

Aberley D (1994) Futures by design: the practice of ecological planning. New Society, Gabriola Island

Alexander C (2001) The nature of order. The Center for Environmental Structure, Berkeley

Bartuska T (1981) Values, architecture and context: the emergence of an ecological approach to architecture and the built environment. In: ACSA annual conference proceedings, San Francisco, Mar 1981

Bartuska T, Young G (1994) The built environment definition and scope. In: Bartuska TJ, Young GL (eds) The built environment: a creative inquiry into design and planning. Crisp, Menlo Park

Beatley T, Manning K (1997) The ecology of place: planning for environment, economy, and community. Island Press, Washington, DC

Birkeland J (2010) Positive development: from vicious circles to virtuous cycles through built environment design. Earthscan, London

Bossel H (2001) Assessing viability and sustainability: a systems-based approach for deriving comprehensive indicator sets. Conserv Ecol 5(2):12. (Online) URL: http://www.consecol.org/vol5/iss2/art12/

Charles J, Kibert J (1999) Reshaping the built environment: ecology, ethics, and economics. Island Press, Washington, DC

Cowan S (2004) Evaluating wholeness. Resurgence 225:56

Crowe N (1997) Nature and the idea of a man-made world: an investigation into the evolutionary roots of form and order in the built environment. MIT Press, Cambridge, MA

Cunningham S (2008) reWealth! McGraw Hill, New York

Edwards A (2010) Thriving beyond sustainability: pathways to a resilient society. New Society, Gabriola Island

France RL (ed) (2008) Handbook of regenerative landscape design. CRC Press, Boca Raton

Franklin C (1997) Fostering living landscapes. In: Thompson G, Steiner F (eds) Ecological design and planning. Wiley, New York

Golley FB (1993) A history of the ecosystem concept in ecology: more than the sum of the parts. Yale University Press, New Haven

Graham R (1993) Restorative design: an interview with Bob Berkebile. Designing a Sustainable Future, in context 35, p 9

Gross M (2010) Ignorance and surprise: science, society, and ecological design. MIT Press, Cambridge

Haggard B (2001) The next step: transforming the building industry to model nature. Hope Dance Nov/Dec 2001

Hawley AH (1950) Human ecology: a theory of community structure. Ronald, New York

Holling CS (1994) New science and new investments for a sustainable biosphere. In: Jansson A (ed) Investing in natural capital: the ecological economics approach to sustainability. Island, Washington, pp 57–97

Holmgren D (2002) Permaculture: principles and pathways beyond sustainability. Holmgren Design Services, Hepburn

International Federation of Landscape Architects (2003) Definition of the profession of landscape architecture. IFLA News, No. 48

Kingsland SE (2005) The evolution of American ecology, 1890–2000. Johns Hopkins University Press, Baltimore

Lemons J, Westra L, Goodland R (eds) (1998) Ecological sustainability and integrity: concepts and approaches. Kluwer, Dordrecht

Likens GE (1992) The ecosystem approach: its use and abuse. Ecology Institute, Oldendorf/Luhe

Mang P (2002) Evolving the new role of development: healing places to heal a planet. Regenesis Collaborative Development Group, Santa Fe. http://www.regenesisgroup.com/articles.php. Accessed 18 Oct 2011

Mang P (2010) Tapping the power of place: stories, cities, and sustainability. Sustainable Santa Fe

McDaniel CN (2006) Design on the edge review. Ann Earth XXIV(3). http://docs.google.com/viewer?a=v&q=cache:UWlG1Y4-I-4J:homepages.rpi.edu/∼mcdanc/opEdsAndCommentary/OrrDesignOnEdgeAnnalsPublished.pdf+McDaniel+design+on+the+edge&hl=en&gl=us&pid=bl&srcid=ADGEESjHnEXiHa3cb5RUjKNPOy8VMLejauMjSBdsBFF1-4tIbMLaTle2nl1dRTOGNGrccAeno516J2ThGMhEFvHR4Jf62blSmA_ZZzTOnRkea0J2QENYjOJfKgI1fb7bkun5jsLZJHMl&sig=AHIEtbR7aGogSklQAB0h7uZbSGzBGJYVrg. Accessed 18 Oct 2011

McIntosh R (1985) The background of ecology: concept and theory. Cambridge University Press, New York

McMurry A (2006) Community health and wellness: a socio-ecological approach. Elsevier, Chatswood

Melby P, Cathcart T (2002) Regenerative design techniques: practical applications in landscape design. Wiley, Hoboken

Murphy T, Reed B (2003) Brattleboro food co-op: preparing the ground for a regenerative market and marketplace:

preliminary report prepared for the Brattleboro Food Co-op. Unpublished regenesis report. http://www.regenesisgroup.com/articles.php. Accessed 18 Oct 2011

Nabhan G (1997) Cultures of habitat. Counterpoint, Washington, DC

Naess A (1989) Ecology, community and lifestyle: outline of an ecosophy. Cambridge University Press, Cambridge

Newman P, Jennings I (2009) Cities as sustainable ecosystems: principles and practices. Island Press, London

Orr D (2002) The nature of design: ecology, culture, and human intention. Oxford University Press, New York

Orr D (2006) Design on the edge: the making of a high-performance building. MIT Press, Cambridge

Reed B (2006) Shifting our mental model – "sustainability" to regeneration. Paper submitted for the conference: rethinking sustainable construction 2006: next generation green buildings, Sarasota, 19–21 Sept 2006

Reed B (2007) Shifting from 'sustainability' to regeneration. Build Res Inf 35:674–680

Regenesis Collaborative Development Group (2006) Bibliography: introduction to thinking behind regenerative design/development. http://www.regenesisgroup.com/pdf/Regenesis_Bibliography.pdf. Accessed 18 Oct 2011

Todd NJ, Todd J (1993) From eco-cities to living machines: principles of ecological design. North Atlantic Books, Berkeley

Wann D (1996) Deep design. Island Press, Washington, DC

Zari MP (2007) Biomimetic approaches to architectural design for increased sustainability. Paper number: 033, School of Architecture, Victoria University, Wellington

Zimmerman M (2004) Being nature's mind: indigenous ways of knowing and planetary consciousness. ReVision 26(4):8, Heldref Publications/Gale Group, Farmington Hills

Regional Air Quality

Erika von Schneidemesser, Paul S. Monks
Department of Chemistry, University of Leicester,
Leicester, UK

Article Outline

Glossary
Definition of the Subject and Its Importance
Introduction
Air Quality Measurements
Health Effects, Ecosystem Effects, and Air Quality Modeling
Current Air Quality Legislation and Metrics
Regional Context
Future Directions
Bibliography

Glossary

Anthropogenic Caused by humans; man-made.

Biogenic Produced by biological processes.

Cohort study A long-term study of the same group of people (the cohort) over time, typically used in medicine social science, and ecology.

Diurnal Relating to the 24 h/day time period.

Dry deposition The deposition of gaseous or particulate species from the atmosphere to a surface without the involvement of precipitation.

Exceedance The amount by which something (in this case an air quality pollutant) exceeds the permissible standard or limit value.

Feedback An internal climate process that amplifies or dampens the climate response to an initial forcing, e.g., the release of methane into the atmosphere due to warmer temperatures, the atmospheric methane increase furthers warming, which causes the release of more methane into the atmosphere, etc.

Megacity City with a population of greater than ten million inhabitants.

Metrics A set of measurements used to quantitatively gauge efficiency, progress, or performance. In this case, as applied to air quality.

Mobile sources Any object that releases pollution that is not stationary, e.g., vehicles, planes, trains, ships, etc.

Radiative forcing A change imposed on the Earth's radiation balance in units of heat flux as watts per square meter (W m^{-2}).

Smog A term derived from smoke and fog applied to extensive contamination by aerosols.

Stationary sources Any object that releases pollution from one set location, e.g., power plants, industrial factories, etc.

Temperature inversion The temperature of the troposphere normally decreases with increasing altitude; however, a temperature inversion results when warm air lies above cool air, which is an extremely stable condition, also known as an inversion layer.

Temporal Pertaining to time.

Wet deposition The deposition of gaseous or particulate species from the atmosphere to a surface during precipitation events (e.g., rain, snow, fog, dew).

Definition of the Subject and Its Importance

What is Regional Air Quality?

- Clean air is considered to be a basic requirement of human health and well-being. However, air pollution continues to pose a significant threat to health worldwide [1].

Air pollution can be defined as "when gases or aerosol particles emitted anthropogenically build up in concentrations sufficiently high to cause direct or indirect damage to plants, animals, other life forms, ecosystems, structures, or works of art" [2]. The state of air pollution is often expressed as air quality (AQ). Air quality is a measure of the concentrations of gaseous pollutants and size, number or mass of particulate matter (PM). Air quality is typically measured using in situ monitors that monitor species such as ozone (O_3), particulate matter, nitrogen oxides, and/or sulfur dioxide.

Why is Regional Air Quality Important?

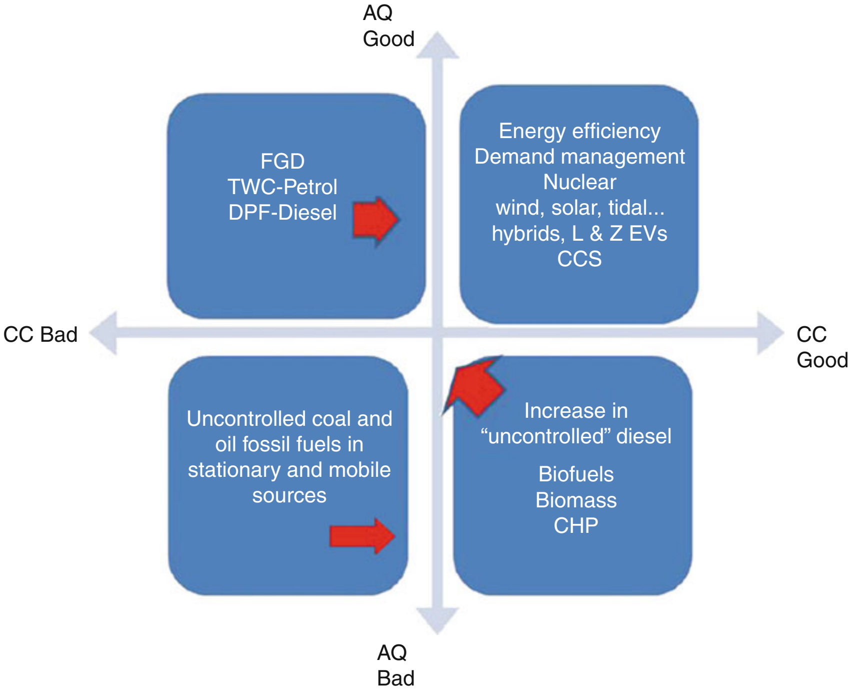
An awareness of the detrimental effects of air pollution came to the forefront because of the catastrophic effects of the "London Smog" episode of 1952 where more than 4,000 people died [3]. Shortly after this, the Clean Air Act was passed by the UK Parliament in 1956 to limit smoke/sulfur dioxide emissions and improve air quality in the cities. Air quality legislation evolved from there, mainly in Europe and the USA, to what it is today. Current air quality legislation is discussed in more detail in section Current Air Quality Legislation and Metrics. Air quality remains an important topic because of its impact on human health, ecosystems, and its interactions with climate change (CC).

Introduction

The Development of Regional Air Quality Monitoring and Legislation

The state of air pollution is often expressed as air quality (AQ). Air quality is a measure of the concentrations of gaseous and aerosol pollutants in the atmosphere. Human health, ecosystems, national heritage (e.g., buildings, monuments), and regional climate are all impacted by air pollution. The implications for human health and ecosystem impacts can range from superficial surface discoloration of buildings to serious adverse health effects, such as loss of months to years of life. These effects are discussed in further detail in section Health Effects, Ecosystem Effects, and Air Quality. Air pollution and air quality control measures have implications for climate change, owing to feedbacks between composition and climate as well as wider atmospheric interactions. Policies to improve air quality may not always have positive effects on climate change and vice versa. Therefore, as we move forward with air quality and climate regulation, interactions and feedbacks from air quality and climate change trade-offs need to be taken into consideration. Some of these trade-offs are shown in Fig. 1. The topic as a whole is addressed in section Regional Context.

Air pollution has a history far longer than the development of contemporary air quality policies. Ozone, a current air quality problem and significant contributor to urban pollution, was discovered in 1839 by a German chemist, Christian Fredrich Schönbein. Air pollution from copper smelting from as far back as Roman times has been detected in ice-core measurements [2]. The plentiful use of fossil fuels in industry and home heating starting in the Middle Ages, latterly in trains, was the main progenitor of degraded air quality through to modern times. A number of extreme air pollution events starting in the late 1800s provided the motivation for modern air quality regulation. A number of smog episodes in London from 1878 to 1962 resulted in excess deaths of hundreds to thousands of people, with the most extreme and well known of these events being the "Big Smoke" of early December 1952 that killed upward of 4,000 people [2–4]. Similar episodes were also recorded in the Meuse Valley, Belgium in 1930, and Donora, Pennsylvania, USA in 1948 where smog events resulted in tens of excess deaths and the illness of thousands of residents [2]. Around the same time, Los Angeles was suffering from a different type of smog, this one photochemical in origin, that arose from the mixture of ozone, nitrogen oxides, and volatile organic compounds (VOCs) produced from a combination of large amounts of vehicle traffic, sunny weather, and persistent temperature inversions [4].



Regional Air Quality. Figure 1
The synergies and trade-offs between policies to improve air quality (AQ) and to reduce greenhouse gas emissions/ improve climate change (CC) (Monks et al. [36])

These "smog" events, among others, spurred the development of air quality regulation in the USA and Europe. The US Air Pollution Control Act was first passed in 1955, identifying air pollution as a national problem. Further air pollution legislation followed with the Clean Air Act in 1963 which set emission standards for stationary sources (e.g., power plants). This was followed by a number of amendments in the 1960s that set standards for motor vehicle emissions. In 1970, the Clean Air Act was passed, which established the National Ambient Air Quality Standards (NAAQS), set new limits on stationary and mobile sources emissions, and allowed states to set their own stricter standards [4]. Catalytic convertors for vehicles in the USA were introduced in 1975, developed in response to the 1970 Clean Air Act, significantly reducing hydrocarbon (HC) and carbon monoxide emissions. Further incarnations of the Clean Air Act and amendments were passed in the years since then, reducing standards for stationary and mobile sources, and mandating control technology. In the USA, California has frequently been ahead of the curve with more stringent air quality standards and tighter standards for vehicles emissions. In 2002, California passed legislation that required automakers to reduce greenhouse gas emissions from vehicles [4].

Just after the first US Air Pollution Control Act, the UK Clean Air Act was passed in 1956, setting controls for household and industrial emissions of smoke. In 1968, the UK Clean Air Act was passed which required tall smoke stacks for industries using fossil fuels to improve dispersal of emissions and avoid local deposition [2]. As of 1972, the UK joined the EU and the majority of air quality policy has been determined by EC directives applicable across Europe. In 1979, the International Convention on Long-Range Transboundary Pollution was adopted to limit emissions of acidifying pollutants. The first directive on ambient air quality was adopted in 1980 and assigned limits to sulfur dioxide and particles [3]. Subsequent directives in 1982 and 1985 put limits on ambient air concentrations of lead and nitrogen dioxide, and in 1989 a directive required all new cars in the EU to run on unleaded fuel [3]. A 1993 directive required the use of catalytic convertors on all new petrol vehicles. In 1996, an EC directive was established to provide a new framework to control levels of sulfur dioxide, nitrogen dioxide, particulate matter, lead, ozone, benzene, carbon monoxide, and hydrocarbons. This included establishing common methods and criteria for assessment of air quality, and setting concentration limits for the pollutants. These limits were then tightened by a following directive in 1999. Current legislation and regulations are outlined in more detail in section Current Air Quality Legislation and Metrics.

Air Quality Pollutants

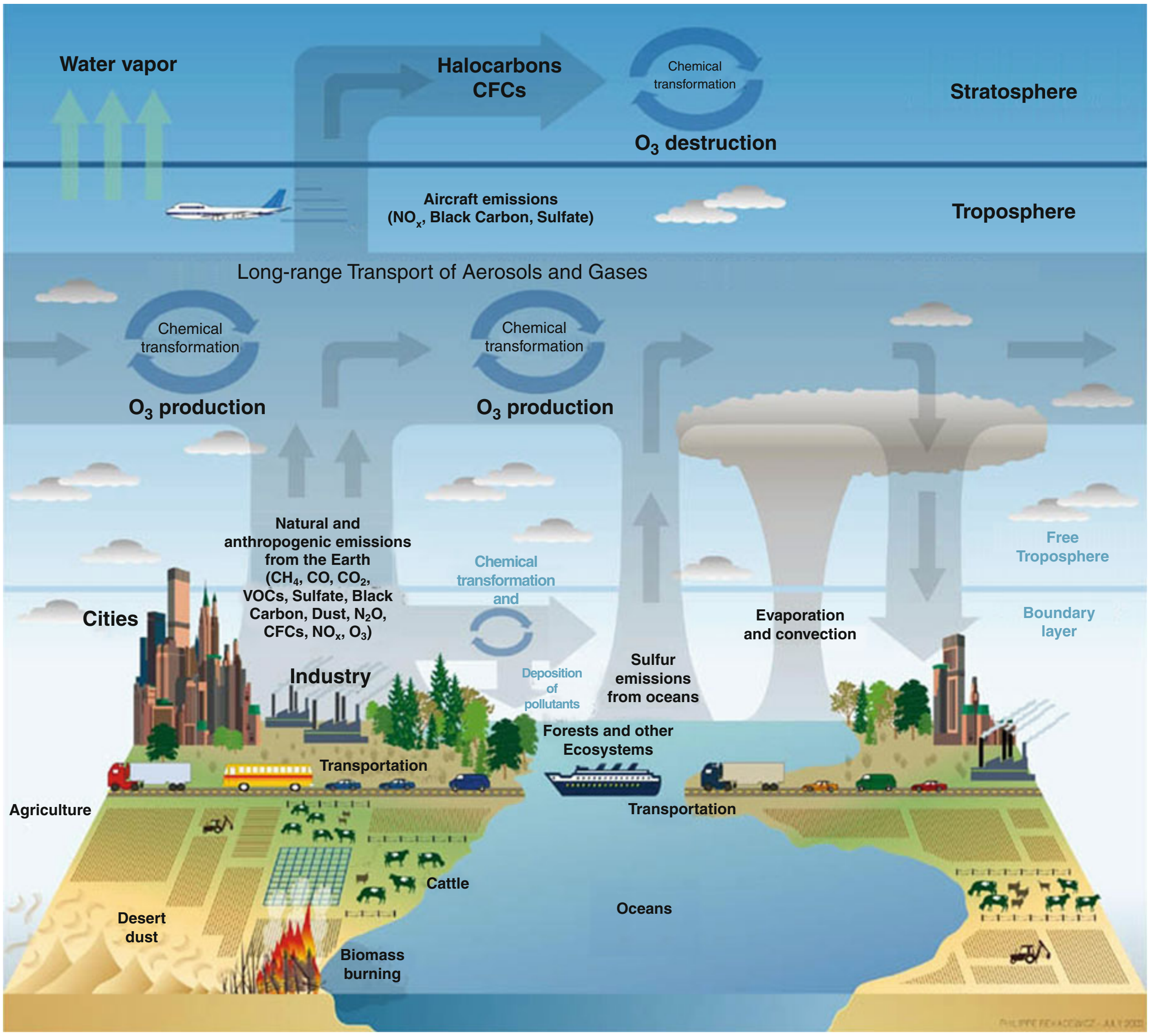
Air pollution has many different components that differ in concentration depending on the superimposition of local, regional, and global emission sources. These emissions are frequently separated into anthropogenic ("man-made") and biogenic ("natural") sources. The components of air pollution can be roughly separated into gaseous and particulate (or aerosol) fractions. Within both gas and particulate pollution there can be primary and secondary pollutants, organic and inorganic compounds. Primary pollutants are those emitted directly into the atmosphere from a source, such as sulfur dioxide or black carbon (BC) (a significant light absorbing component of soot). Secondary pollutants are those formed in the atmosphere from chemical and/or physical transformations of primary pollutants, such as ozone (O_3) and secondary organic aerosol (SOA).

Regional air pollution results from the combination of emissions, which vary both spatially and temporally and meteorological factors (that allow the pollutants to build up in an area or transport pollutants into an area where they can mix with local emissions) causing poor regional air quality. Figure 2 shows different emissions sources and meteorological components that have an impact on regional air quality. In addition to wind transporting pollutants, sunlight plays an important role in photochemical reactions in the atmosphere that create some of the secondary pollutants. Air pollutants are then removed from the atmosphere through dry deposition (e.g., settling) or wet deposition (e.g., rain scavenging). The following two sections will discuss the two main components of air pollution – gaseous pollutants and particulate pollutants in terms of their sources, sinks, atmospheric transport, and transformations.

Gaseous Pollutants

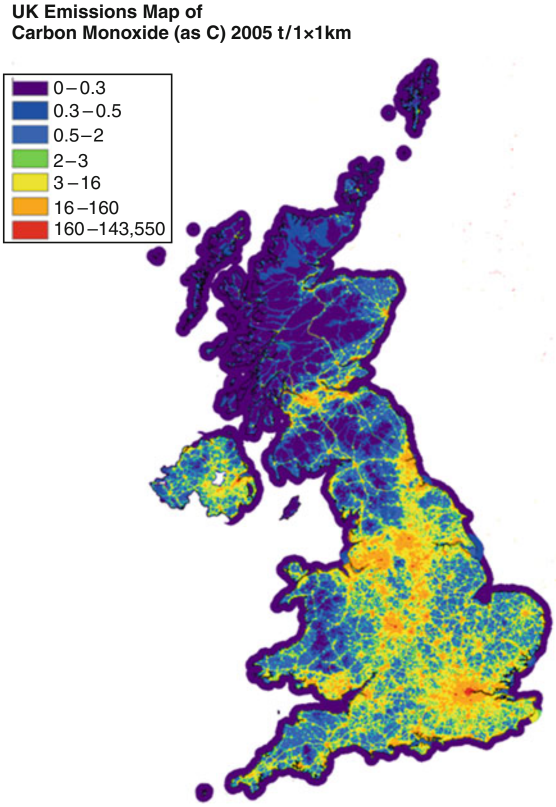
There are myriad different gaseous species that are emitted into the atmosphere from various air pollution sources. Some of the major gaseous pollutants are carbon monoxide (CO), nitrogen compounds (e.g., NO, NO_2, HNO_3), sulfur compounds (e.g., SO_2), hydrocarbons (HCs) including methane and nonmethane hydrocarbons (NMHCs), and photochemical oxidants (e.g., O_3). Other gaseous pollutants, such as carbon dioxide (CO_2) and methane (CH_4), will not be addressed here because they are not a major focus for regional air quality concerns because their longer lifetimes qualify them as global pollutants.

Carbon monoxide's primary source is from incomplete combustion. In addition to being an important primary pollutant, it is also an important precursor compound for the formation of ozone, as well as a secondary pollutant itself, formed from the oxidation of methane and other NMHCs by the OH radical [5]. It has a lifetime of a couple months which makes it a regional and global scale pollutant that can be transported significant distances from its emission source. Carbon monoxide is primarily removed from the atmosphere by reaction with OH, a small amount being removed by deposition. Ambient levels of 0.15–10 ppmV are common in urban areas mainly owing to road transport–related sources [6, 7]. In the USA, mobile sources (including non-road mobile sources) make up 80% of national CO emissions, while in the UK 47% of CO emissions are attributed to road transport sources. Figure 3 shows how CO emissions largely follow the road network in the UK, with the highest emissions in urban areas [8]. Furthermore, the high correlation of CO with population density in Asia is shown in Fig. 4 [9]. Carbon monoxide emissions have shown significant reductions over the past 2 decades [7]. These emission reductions are largely credited to the increased use of catalytic convertors in cars. Total CO emissions in the USA have decreased 68% since the 1990s until 2008, and currently all monitoring stations in the USA show that no areas are in nonattainment for the 8-h CO standard of 9 ppmV [10]. Similarly, CO emissions in the UK have been reduced by 71% from 1990 to 2005 in the UK. Elevated levels of CO observed in rural/remote areas are possible because of agricultural or biomass burning and forest fires.



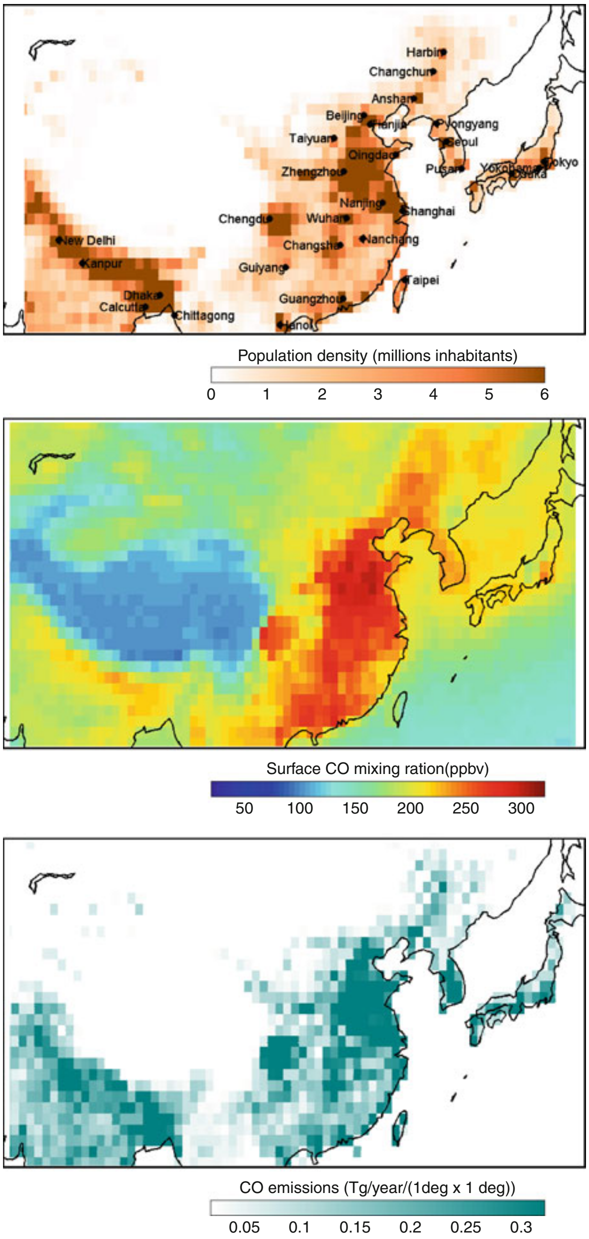
Regional Air Quality. Figure 2
Summary of emissions, emissions sources, and atmospheric processing of air pollutants

Sulfur dioxide emissions originate primarily from coal-fired power plants. Once in the atmosphere SO$_2$ is oxidized to sulfuric acid (H$_2$SO$_4$) both homogeneously and heterogeneously in the liquid and gas phases [11]. The resultant sulfuric acid can be deposited in the gas phase or can condense with water vapor onto aerosol particles or cloud drops to produce aqueous phase sulfuric acid which can then be deposited through rain or fog. Acid deposition came to the forefront of environmental issues in the 1950s and 1960s when the acidification of Scandinavian lakes was linked to sulfur emissions in Europe. A Swedish study in 1972 connected sulfur dioxide emissions with negative environmental impacts which led to an international effort to reduce acidification [2, 3]. It was later also determined that emissions other than SO$_2$ were also contributing to the acid deposition, such as direct emissions of HCl and nitric acid formed from nitrogen oxides emission and chemical transformation. Efforts to reduce acid deposition spawned international agreements such as the Convention on Long-Range Transboundary Air Pollution (LRTAP), which was the first such agreement to deal with international air pollution issues. Measures in the USA to reduce and



Regional Air Quality. Figure 3
Spatially disaggregated emissions of carbon monoxide in the UK from 2008 (Murrells et al. [8])

monitor sulfur emissions included amendments to the Clean Air Act of 1970, as well as the creation of the National Atmospheric Deposition Program of 1977 [2]. Significant reductions in SO_2 emissions were achieved in North America and Europe, resulting in some regeneration and recovery of many forests and water bodies in these areas. The current guideline set by the World Health Organization (WHO) for SO_2 is 20 μg m^{-3} (24-h average). A study looking at annual average SO_2 concentrations for the late 1990s of various megacities (cities with populations greater than ten million) worldwide found that a number of cities, such as Dhaka, Bangladesh, Beijing, China, and Shanghai, China were still above this guideline; most other cited megacities in the study had annual average concentrations of just under 25 μg m^{-3} [12]. Additionally, SO_2 emissions from megacities in Asia accounted for 30% of the ambient concentrations measured



Regional Air Quality. Figure 4
(*top*) Population density over China and surroundings. All cities with more than two million inhabitants are indicated. (*middle*) MOPITT carbon monoxide mixing ratios at the surface level; average measurements from March 2000 to June 2007. (*bottom*) The EDGAR carbon monoxide emission inventory for the year 2000. All data are on a 1° × 1° grid (Clerbaux et al. [9])

regionally, including areas hundreds of kilometers from the urban centers [13]. Levels of SO_2 measured in the USA declined by 59% from 1990 to 2008, and are currently well below the 0.03 ppmV ($\sim$80 μg m^{-3}) national standard. The majority of the monitoring sites in the USA report SO_2 annual average levels between approximately 0.002 and 0.008 ppmV for 2008 [10].

Similar to SO_2, power plants are a significant emissions source for nitrogen oxides (NO_x). However, NO_x has many other important emission sources including mobile sources, which make up the majority of global NO_x emissions, as well as biomass burning, and natural emissions from soils and lightning [2, 5]. NO_x is the collective term for NO and NO_2 as the photolysis of NO_2 to NO and the subsequent regeneration of NO_2 *via* reaction of NO with ozone is sufficiently fast, in the moderately polluted environment, for these species to be in dynamic equilibrium, namely,

$$NO_2 + h\nu \rightarrow NO + O \quad (1)$$

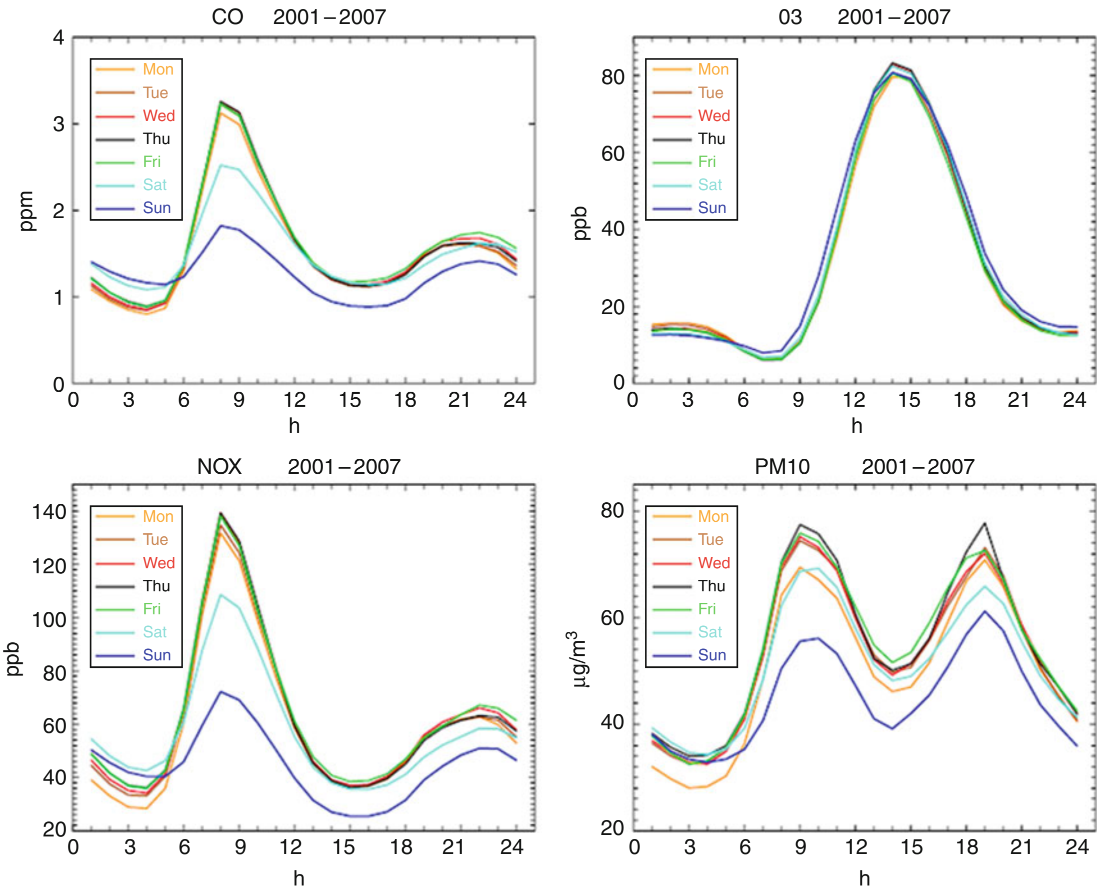
$$O + O_2 + M \rightarrow O_3 + M \quad (2)$$

$$O_3 + NO \rightarrow NO_2 + O_2 \quad (3)$$

Therefore, at suitable concentrations, ambient NO, NO_2, and O_3 can be said to be in a photochemical steady state or photostationary state (PSS) [14], provided that they are isolated from local sources of NO_x and that sunlight intensity is relatively constant.

The majority of NO_x emissions are initially emitted as NO. For example, the majority of NO_x from power stations is released from the stack as NO. Also, a study of vehicle emissions in Europe (for Euro-II emission standards) showed that for passenger cars, only 5% of NO_x is emitted as NO_2 on average [15]. Since the Euro-II emission standards were introduced NO_x standards have been reduced through the introduction of Euro-III, IV, and V. This has resulted in an increase in the fraction of NO_x emitted as primary NO_2, likely due in large part to particle filters fit to diesel vehicles [16]. In urban environments, NO_x emissions typically show diurnal cycles, peaking during rush hours, owing to the dominant source from the transportation sector. The data in Fig. 5 show typical source time profiles from Mexico City [17]. The adverse health effects of NO_x stem from NO_2 which can be inhaled and passed into the pulmonary region causing tissue damage; NO is believed to be largely nontoxic at ambient levels [18]. The greater impact, in terms of health effects and air quality, results from the role that NO_x plays in ozone formation. NO_x along with volatile organic compounds (VOCs) and CO are the important precursor compounds for the formation of ozone (see later).

The main anthropogenic sources of VOCs are from mobile sources, evaporation from liquid fuels, and industrial sources. VOCs such as isoprene and monoterpenes also have a significant biogenic source from vegetation [19]. VOCs are defined as nonmethane hydrocarbons plus heavy hydrocarbons plus carbonyls plus halocarbons, typically $<C_{20}$ [20]. VOCs include alkanes, alkenes, alkynes, and aromatics, as well as oxygenated organics that have low boiling points (50–260°C). Certain VOCs such as the highly reactive "BTEX" compounds (benzene, toluene, ethylbenzene, and xylenes) and 1,3-butadiene that are often prevalent at significant concentrations are frequently monitored even if other VOCs are not [21–24]. Highly reactive aromatics, as well as some other VOCs, are important precursor compounds to secondary organic aerosol (SOA) formation (see discussion in the particulate matter/aerosol pollutants section). Methane is also a VOC, but those discussed here are limited to nonmethane VOCs (NMVOCs). Methane has very different sources and because of its lifetime is a global pollutant; it also has very important effects on climate. A study of NMHCs in 28 US cities reported mean mixing ratios of individual HCs of 9 pptV–8740 pptV, with Los Angeles typically reporting among the highest mixing ratios [21]. A similar study measuring NMHCs in 43 cities in China reported a range of minimum and maximum mixing ratios from 40 pptV to 58,300 pptV for the same set of compounds [25]. On the other end of the spectrum, measurements at a remote, high altitude observatory at Izana, Tenerife, reported mean values for a similar set of individual NMHCs ranging from 1 pptV to 501 pptV [26]. A comparison of NMHC (and CO) measurements from global cities found that, in agreement with the small sampling of measurements discussed here, the highest mixing ratios were found in Asian and Latin American cities [7]. While ambient air quality standards do not include VOCs, many vehicular emissions standards do, such as the European emissions standards for passenger cars or the US "Tier" regulations. The standards set for VOCs are formulated

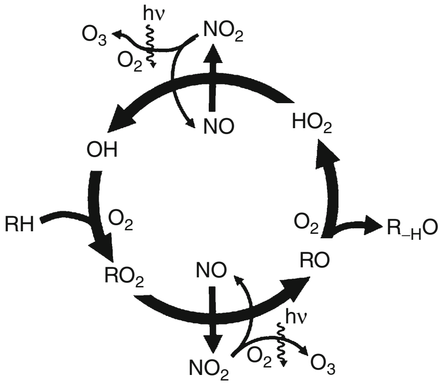


Regional Air Quality. Figure 5

Diurnal cycle of carbon monoxide, nitrogen oxides, ozone, and PM_{10} in Mexico City, averaged for 39 monitoring stations throughout the city over the 2001–2007 period (Stephens et al. [17])

to achieve ozone standards, as the many of the VOCs themselves pose relatively low health risks owing to minimal exposures. Some VOCs, such as 1,3-butadiene and benzene are genotoxic carcinogens and have significant health risks for any amount of exposure. These compounds along with some semi-volatile organic compounds, such as polycyclic aromatic hydrocarbons (PAHs) frequently have specific air quality standards because of their carcinogenic properties [2, 18, 24].

Ozone is currently a problematic regional air quality issue for many locations. Ozone is formed in the atmosphere from the photochemically driven reactions of NO_x, VOCs, and CO. The atmospheric cycle of reactions that form ozone are shown in Fig. 6, where RH represents the VOCs/hydrocarbons. High levels of ozone can be found in urban areas, as well as in rural



Regional Air Quality. Figure 6

Atmospheric cycle for the production of ozone. RH are hydrocarbons/non-oxygenated volatile organic compounds (VOCs) (Reproduced from Fowler et al. [32])

areas. Ozone is a secondary pollutant and takes time to form in the atmosphere, during that time can be transported away from the emission source. The time necessary for ozone formation depends on factors such as sunlight and the availability of the precursor compounds, and is typically on the order of hours. Modern day average background O_3 mixing ratios in the Northern Hemisphere range from 20 to 45 ppbV ($\sim$39–88 μg m^{-3}). Background in this case is defined as the fraction of ozone that is not attributed to anthropogenic sources of local origin and is therefore resulting from stratospheric ozone transport to the troposphere, ozone production from natural sources of NO_x and VOCs (e.g., lightning, biogenic emissions), or long-range transport of ozone from distant pollutant sources [27]. The contribution of background ozone, specifically that from long-term transport, is a concern for many locations that are in exceedance of air quality standards, as the background contribution can make the difference between being in attainment and not. This is of particular concern in the Western US where influence from Asian emissions may be making it more difficult to achieve current regulatory air quality standards [28]. Many urban areas, beyond those in the Western US, are often in exceedance of their respective air quality standards for ozone, such as cities in the USA, Canada, Europe, and China [29–34]. These exceedances often affect rural areas as well [35]. Ozone has adverse impacts on human health, as well as vegetation. Surface ozone is associated with respiratory problems and premature human mortality. In addition, it can damage plants, reduce photosynthesis and growth, leading to reduced crop yields [36].

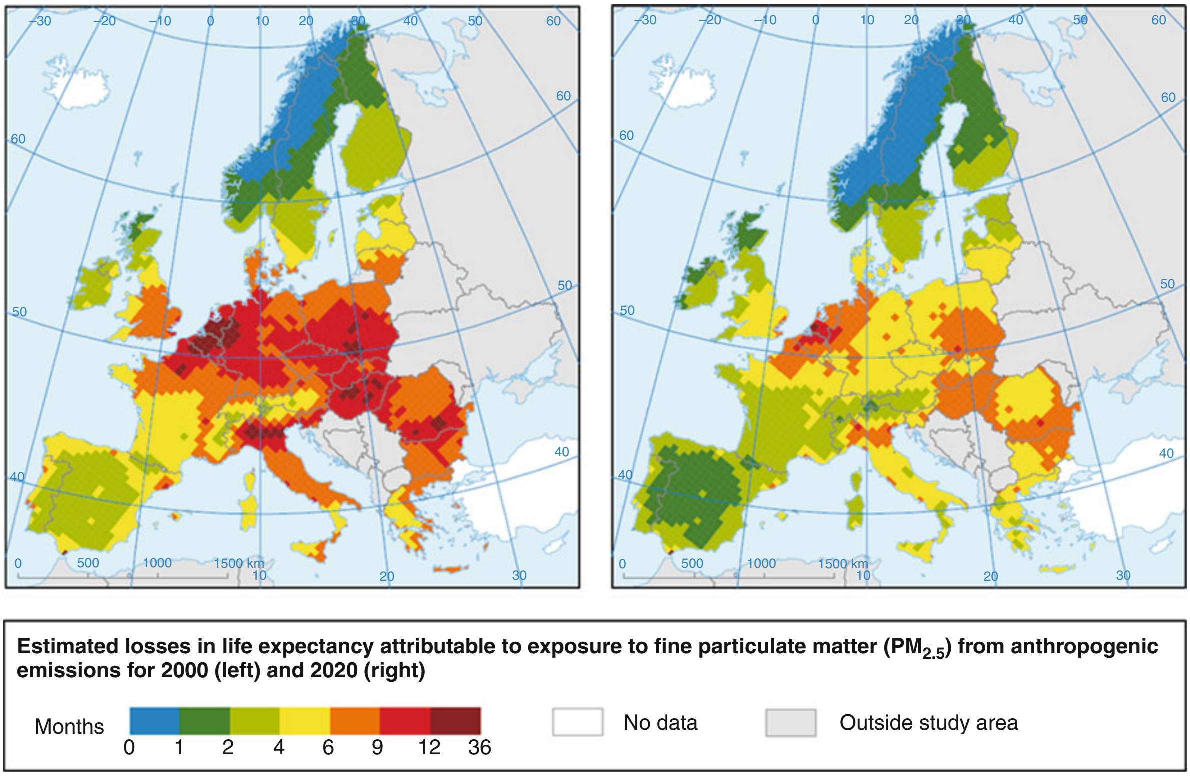
Many of the pollutants discussed above, including carbon monoxide, sulfur oxides, nitrogen oxides, and ozone, as well as particulate matter (see below), are classified as “criteria pollutants” by the US EPA. These criteria pollutants are among the most common air pollutants and are used as indicators of air quality [37].

Particulate Matter/Aerosol Pollutants

Particulate matter is an air quality concern because of its adverse health effects, as well as its contribution to reductions in visibility. Particulate matter is currently monitored, in the air quality area, based on size-resolved fractions. These are typically PM_{10} (particles with an aerodynamic diameter of 10 μm or smaller) and $PM_{2.5}$ (particles with an aerodynamic diameter of 2.5 μm or smaller). This size fractionation is also typically referred to as the coarse mode (the fraction between PM_{10} and $PM_{2.5}$) and the fine mode (equivalent to $PM_{2.5}$). Particles in the fine mode are a more significant health concern than those in the coarse mode because of their ability to penetrate deeper into the lungs than particles in the coarse mode, remain in the air longer, transport over longer distances, and penetrate more easily into indoor environments [38]. The scale of the possible impact of adverse health effects from particulate matter is shown in Fig. 7 which depicts the estimated loss of life expectancy attributable to $PM_{2.5}$ from anthropogenic emissions in 2000 and 2020 in Europe [39].

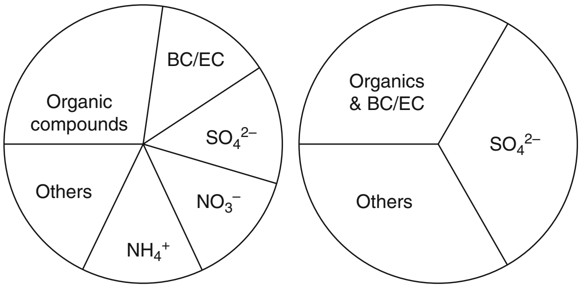
Particulate matter (PM) is composed of many different species, including carbonaceous aerosols, inorganic species (sulfate, nitrate, ammonium, chloride), trace metals, and crustal elements from dust and water [5]. The relative contribution of aerosol components to the total aerosol burden for the present atmosphere, based on a model simulation for 1990 using emission estimates from 1990 and later, is shown in Fig. 8 [40]. Aerosols technically refer to the suspension of fine solid or liquid particles in a gas; however, the term aerosols is more commonly used to refer to just the particulate/nongaseous component, which is how it will be used here [5]. There are anthropogenic and natural sources of particulate matter. Natural sources of PM include volcanic eruptions, sea spray, biological debris, and dust. These types of PM emissions tend to fall into the coarse mode (PM_{10}–$PM_{2.5}$) and are emitted owing to mechanical processes [5]. Anthropogenic particulate matter is emitted primarily from combustion sources and falls largely into the fine mode ($PM_{2.5}$) [38]. In many locations the majority of these anthropogenic emissions stem from fossil fuel combustion use by the transportation sector [18, 41], but other sources include stationary fuel combustion (e.g., power plants), industry, and biomass burning. The schematic representation in Fig. 9 summarizes aerosol sources, sinks, and transformations in the atmosphere, by particle size [5, 42].

In addition to primary aerosols, secondary aerosols, i.e., those formed in the atmosphere by gas-to-particle


-30
-20
-10
0
10
20
30
40
50
60
60
50
40
0
500
1000
1500 km
10
20
30
Estimated losses in life expectancy attributable to exposure to fine particulate matter ($PM_{2.5}$) from anthropogenic emissions for 2000 (left) and 2020 (right)
Months
0
1
2
4
6
9
12
36
No data
Outside study area

Regional Air Quality. Figure 7

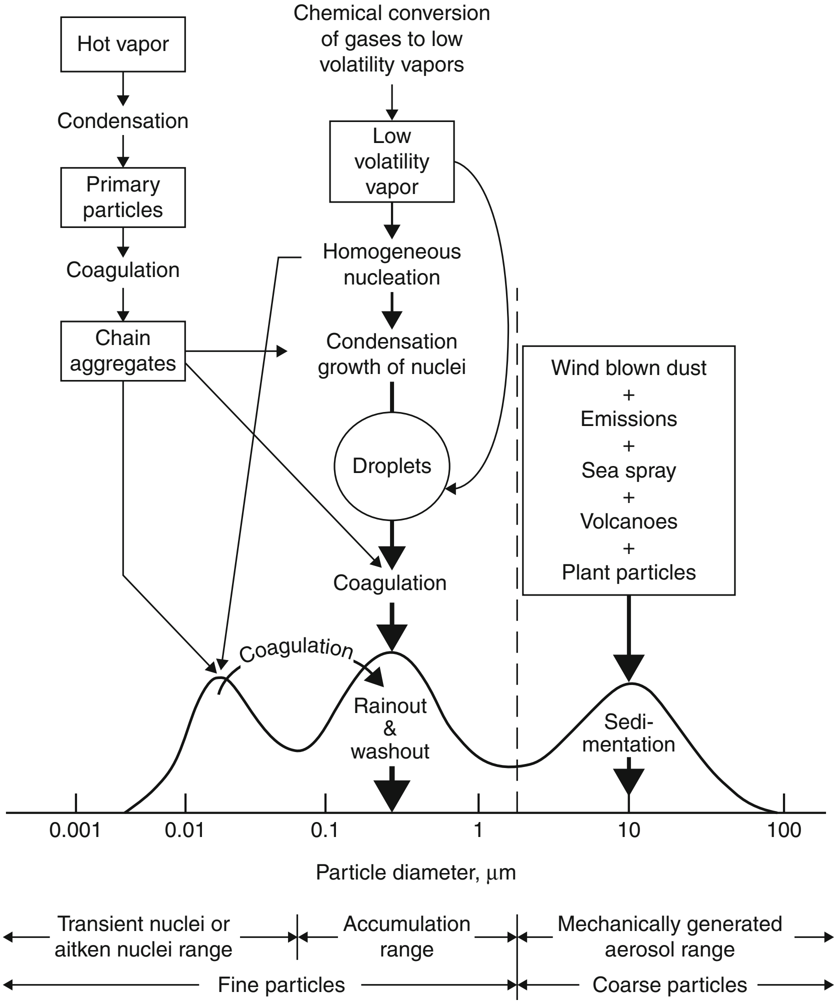
Health impact of particulate matter (PM) mass concentrations (μg m^{-3}). Loss in statistical life expectancy (months) that can be attributed to anthropogenic contributions to $PM_{2.5}$ for the year 2000 (*left*) and for 2020 (*right*) for the Clean Air for Europe (CAFE) baseline scenario, which takes into account changes in air quality standards and reductions in emissions (IIASA; EEA [39])



Regional Air Quality. Figure 8

Relative contribution of aerosol components to the total aerosol burden (28.9 Tg) for the present atmosphere, based on a model simulation from 1990 using emission estimates from 1990 and later. Secondary organic aerosol (SOA) produced from anthropogenic sources (SOAa), secondary organic aerosol from biogenic VOC oxidation (SOAb), primary organic aerosol (POA), black carbon (BC), and methane sulfonic acid (MSA) (Tsigaridis et al. [40])

conversion and/or condensation of gaseous compounds onto preexisting aerosol particles can be a significant fraction of the PM. Secondary aerosols can be organic or inorganic, formed primarily from (photo)chemical reaction of VOC precursor emissions or the oxidation of NO_x and SO_2 [43]. Secondary organic aerosols (SOAs) are formed in a two-step process: (1) the production of the organic aerosol compound from the reaction of parent organic gases, (2) the partitioning of the organic compound between the gas and particulate phases forming SOA [5]. Secondary organic aerosol concentrations tend to be higher during periods of greater photochemical reactivity (e.g., summer months), but can be significant year-round depending on the location, meteorology (e.g., humidity and temperature), solar radiation, the existing aerosols and availability, abundance, and reactivity of gas-phase precursors and oxidants [44–47]. The atmospheric processes



Regional Air Quality. Figure 9
Schematic of the distribution of atmospheric aerosols, showing principal modes, sources, and particle formation and removal mechanisms (Whitby and Cantrell [79])

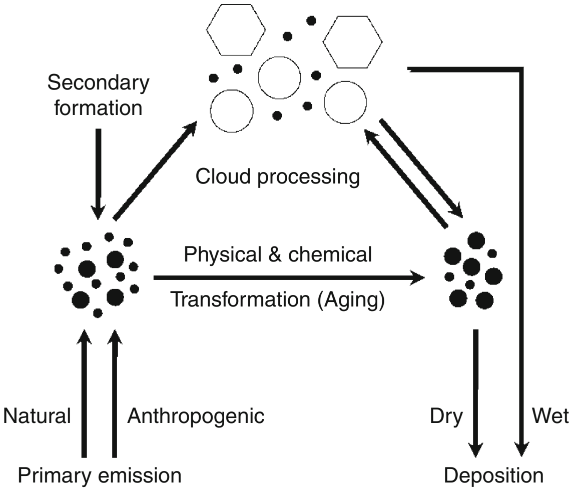
undergone by aerosols after emission or formation in the atmosphere are detailed in Fig. 10 [44, 48].

Concentrations of reported PM range from less than 1 µg m^{-3} (PM_{10}) in a remote location such as the Antarctic to 340 µg m^{-3} (PM_{10}) and 194 µg m^{-3} ($PM_{2.5}$) annual average concentrations for Lahore, Pakistan in 2007, which is extremely polluted and has some of the highest PM concentrations recorded [49–51]. The WHO annual average guideline for PM_{10} is 20 µg m^{-3} and 10 µg m^{-3} for $PM_{2.5}$. Annual average concentrations in 2001 of $PM_{2.5}$ and PM_{10} from three European cities ranged from 7.8 to 28.0 µg m^{-3} to 15.7–41.0 µg m^{-3}, respectively [52].

Various types of modeling have been established that are able to determine the contributions of different types of aerosol sources to ambient particulate matter samples, including factor analysis and chemical mass balance (CMB) modeling. More information about these types of models can be found in section Modeling.

Air Quality Measurements

Networks of air pollution monitoring stations have been set up across many areas of the globe, especially in developed countries. In the USA, for example,

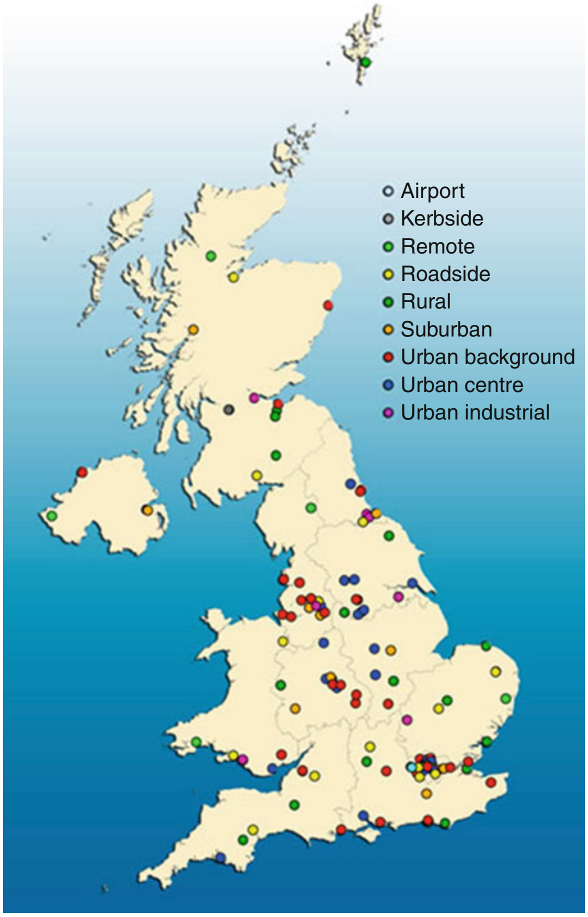


Regional Air Quality. Figure 10
Primary emissions, secondary formation, and atmospheric processing of natural and anthropogenic aerosols (Fuzzi et al. [48]; Poeschl [44]; Monks et al. [36])

ambient air quality monitoring programs are required of states by the US EPA to assess and determine compliance with the national ambient air quality standards. The air quality monitors and monitoring programs must comply with design and quality assurance requirements, as well as using monitoring techniques based on federal reference methods. The results from these monitoring networks are then reported to the US EPA annually [18]. There are different levels of monitoring networks in the USA, including the National Air Monitoring Stations (NAMS), the State and Local Air Monitoring Stations (SLAMS), and the Photochemical Assessment Monitoring Stations (PAMS). The national network is for long-term monitoring to provide a systematic, consistent database for comparison with other air quality data and trend analysis. The state and local network is for determining compliance with air quality standards, while the PAMS network is designed for monitoring ozone in areas of persistently high ozone concentrations [18]. More information on monitoring and the federal reference methods used in the US air quality networks is available on the US EPA Web site and in Godish (2004) [18].

The overarching air monitoring network for all of Europe, which focuses on transboundary air pollution, was established as part of the Convention on Long-Range Transboundary Air Pollution (LRTAP). Within that convention, the EMEP programme (Co-operative Programme for Monitoring and Evaluation of the Long-range Transmission of Air Pollutants in Europe) was tasked to "regularly provide governments and subsidiary bodies under the LRTAP Convention with qualified scientific information to support the development and further evaluation of the international protocols on emission reductions negotiated within the Convention" [53]. At this point, data are collected for the following areas of interest: acidification and eutrophication, ground-level ozone, persistent organic pollutants (POPs), heavy metals, and particulate matter. These are monitored using air and precipitation sampling methods. In addition to measurements (monitoring), EMEP also collects emissions data, and uses modeling to look at atmospheric transport and deposition of air pollutants [53]. The focus here will be on the established monitoring network. There are over 300 sites in the monitoring network in 37 countries, and more than 20 participating institutions. There are three different levels of monitoring stations within the EMEP network depending on the number and type of measurements being conducted, but all EMEP stations are located in areas classified as measuring regional or global background. Many of the EMEP stations are also part of or working in cooperation with other European monitoring framework initiatives, such as the World Meteorological Organization-Global Atmospheric Watch programme (WMO-GAW), national monitoring networks that report to the European Commission under the European Union Air Quality Directives, and other national and local monitoring networks [54]. The European Air Quality Directives focus on regional air pollution.

In addition to the EMEP network, many countries also have monitoring networks that have been established in urban areas to monitor roadside pollution levels, and urban background levels. For example, the UK has an extensive air quality network, across the country (UK Automatic Urban and Rural Network, AURN) and in the major urban areas, such as London (London Air Quality Network, LAQN). The UK AURN has over 100 monitoring stations located across the country for NO_x, $PM_{2.5}$, PM_{10}, SO_2, CO, and/or O_3. A map showing the distribution of monitoring stations and station types, given by monitoring in compliance with AQ legislation, is shown in Fig. 11.



Regional Air Quality. Figure 11
Monitoring stations in the UK, shown by station classification (http://www.airquality.co.uk/monitoring_networks.php?n=aurn)

Health Effects, Ecosystem Effects, and Air Quality

Air pollution was not always associated with negative connotations. Before the Second World War, air pollution (e.g., car exhaust, plumes from smoke stacks) were associated to a certain degree with growth and prosperity [55]. Going back as far as 1873 London was affected by smogs, the most infamous of which was the London smog of December 1952 [18]. The word smog originates from the combination of the words smoke and fog. The smogs of London were caused by extremely high concentrations of smoke and sulfur dioxide emissions largely from coal burning. During the 1940s, it became clear that cities like Los Angeles had severe air quality problems owing to a different kind of smog – a photochemical smog. This type of smog stems from high concentrations of oxidants (e.g., ozone) and peroxide compounds that are produced through photochemical reactions. The basic reaction scheme for the formation of ozone, the primary ingredient in photochemical smog is

$$\mathrm{VOC} + \mathrm{OH} \xrightarrow{[\mathrm{O_2}]} \mathrm{RO_2} + \mathrm{H_2O} \tag{4}$$

$$\mathrm{CO} + \mathrm{OH} \xrightarrow{[\mathrm{O_2}]} \mathrm{HO_2} + \mathrm{CO_2} \tag{5}$$

$$\mathrm{RO_2} + \mathrm{NO} \xrightarrow{[\mathrm{O_2}]} \text{secondary VOC} + \mathrm{H_2O} + \mathrm{NO_2} \tag{6}$$

$$\mathrm{H_2O} + \mathrm{NO} \rightarrow \mathrm{OH} + \mathrm{NO_2} \tag{7}$$

$$\mathrm{NO_2} + h\nu \rightarrow \mathrm{NO} + \mathrm{O} \tag{8}$$

$$\mathrm{O} + \mathrm{O_2} + \mathrm{M} \rightarrow \mathrm{O_3} + \mathrm{M} \tag{9}$$

where RO_2 represents a chain organic with an attached O_2. The OH radical is formed from photolysis of ozone, formaldehyde, and other secondary VOCs [56]. Since these early smogs which documented a connection between exposure to elevated pollution levels and acute illness in the affected population, a number of epidemiological, toxicological, and occupational exposure studies have documented the association between aerosols and adverse health effects [18, 38, 57]. Health effects can range from eye irritation to cardiovascular system related mortality. The Harvard Six Cities study was a cohort study which estimated the effects of air pollution on mortality, while also controlling for other individual risk factors, such as smoking [58]. Over 8,000 adults in six US cities were included in the study that lasted approximately 15 years. Statistically significant associations were found between air pollution and mortality with positive associations between air pollution and death from lung cancer, as well as air pollution and cardiopulmonary disease. The air pollutants most associated with the mortality were fine particulates and sulfate [58]. A later cohort study from 1982 to 1989 using data from 151 US cities and over 550,000 adults found that increased mortality was associated with sulfate and fine-particulate air pollution and that particulate air pollution was linked with cardiopulmonary and lung cancer mortality, very similar to the Six City study findings [59]. Later studies and reanalysis of the data found that there was no consistent toxicological evidence to suggest that long-term exposure to sulfates had significant effects on the cardiovascular system [60]. The association of sulfate with adverse health effects remains under debate. Figure 12 shows life expectancies versus $PM_{2.5}$ concentrations from a study looking at fine-particulate air pollution and life expectancy throughout the USA [61]. Additional studies in Canada, Germany, Finland, and the Czech Republic also reported comparable results for particulate pollution and adverse health effects, such as respiratory symptoms and hospitalizations and mortality [38]. For more detailed information on air pollution and health effects, a number of books and review articles are recommended in the Further Reading section of the Bibliography.

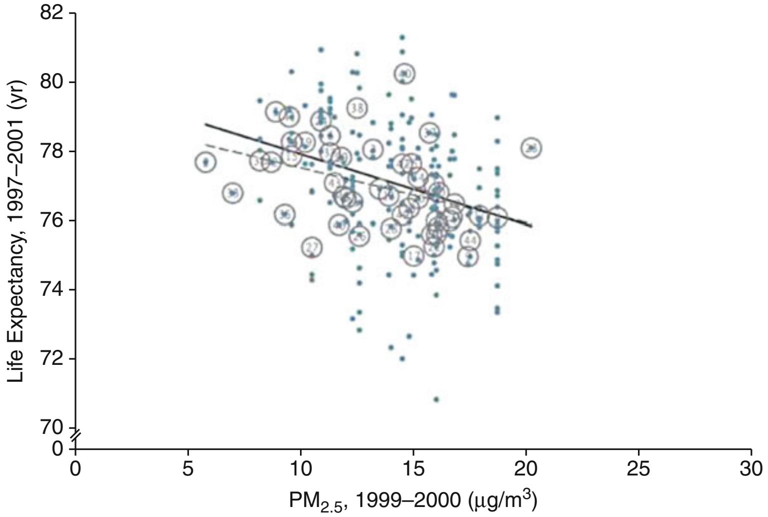
In addition to adverse human health effects, air pollution can damage vegetation, buildings, and cause loss of revenue due to impaired visibility. Vegetation effects include reduction in yields, leaf damage, loss of sensitive species, and subsequent reduction in plant diversity and sensitivity to other environmental stresses [62]. Buildings and other outdoor structures can be damaged by air pollution as well. The blackened walls of historical monuments and other buildings, such as cathedrals, are because of high levels of aerosols depositing on the surface. Physical erosion owing to windblown dust can impact buildings, and some building materials (mainly certain kinds of stone, such as limestone or marble) can also be chemically eroded by acid aerosols [18].

Modeling

The fate of any chemical species (C_i) in the atmosphere can be represented as a continuity or mass balance equation such as

$$\frac{dC_i}{dt} = \frac{duC_i}{dx} - \frac{dvC_i}{dy} - \frac{dwC_i}{dz} + K_z\frac{dC_i}{dz} + P_i - L_i + S_i + \left(\frac{dC_i}{dt}\right)_{\text{clouds}} \tag{10}$$

where t is the time, u, v, and w are the components of the wind vector in x, y, and z accounting for the horizontal and vertical large-scale transport. Small-scale turbulence can be accounted for using K_z, the turbulent diffusion coefficient, P_i and L_i are the chemical production and loss terms, and S_i are the sources owing to emissions. Cloud processes (vertical transport, washout, and aqueous phase chemistry) are represented in

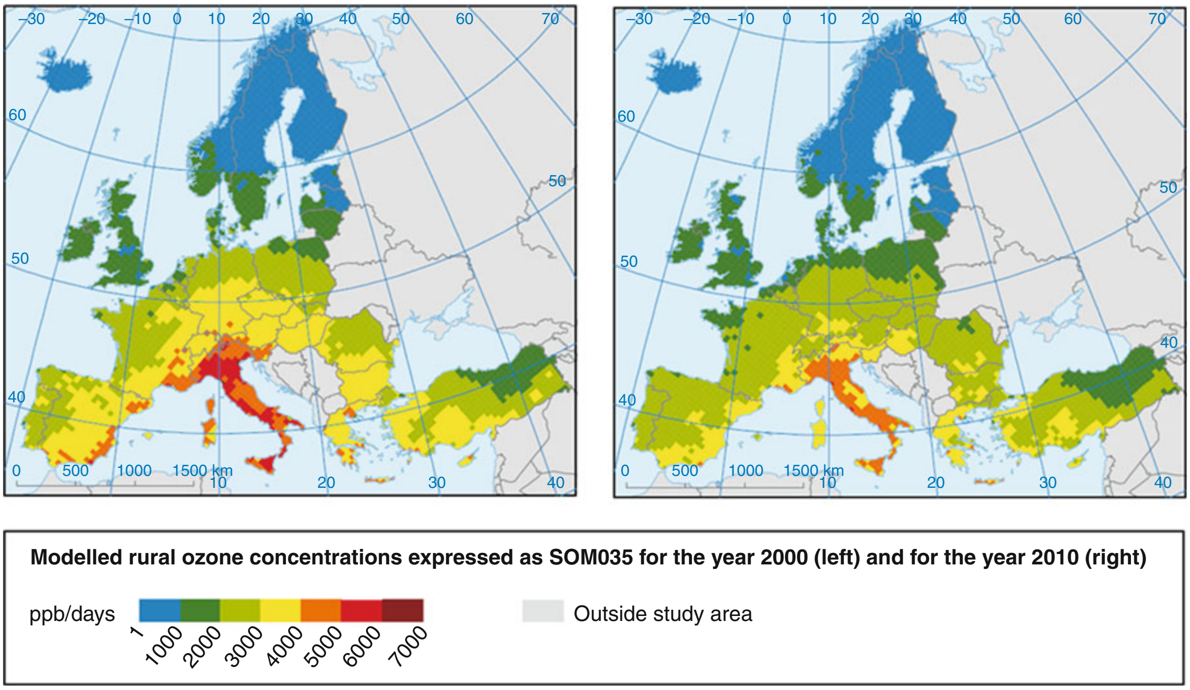


Regional Air Quality. Figure 12
Cross-sectional life expectancies for 1997–2001, plotted against PM2.5 concentrations for 1999–2000 for the USA. Dots and circles labeled with numbers represent population-weighted mean life expectancies at the county level and the metropolitan-area level, respectively. The solid and broken lines represent regression lines with the use of county-level and metropolitan-area-level observations, respectively (Pope et al. [59])

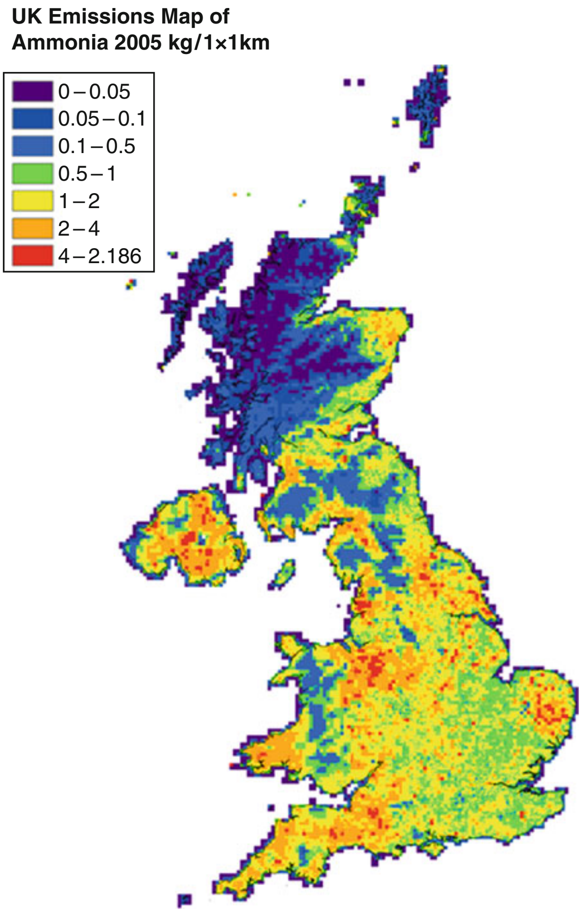
a cloud processing term. The application of this type of equation is the basis of chemical modeling. There are two main types of modeling that deal with air pollution and air quality objectives, atmospheric (chemical) transport models and statistical models [5].

Atmospheric Models

Atmospheric (chemistry) transport models are based on the fundamental description of atmospheric physical and chemical properties [5]. The reason for the "chemical" term in parentheses is that there are a number of models which model only transport and do not necessarily always include chemical transformations, but the general principles behind the models are the same. There are two basic kinds of atmospheric models – Lagrangian and Eulerian. Lagrangian models track the path of a given air parcel (and the concentrations of the chemical components therein) as it is advected in the atmosphere, while Eulerian models describe the concentrations in an array of fixed (nonmoving) computational cells [5]. The concentration of any chemical species in the atmosphere are controlled by four types of processes, namely emissions, chemistry, transport, and deposition (see Eq. 10) that are taken into account in the atmospheric models [19]. Atmospheric models can be run on different scales from local and regional to global. The regional or local models are most applicable to regional air quality and are frequently used as predictors for possible future scenarios, including prediction of air quality, cost-benefit, and environmental impact analyses of pollutant increases or reductions. An example is shown in Fig. 13 where the model has been used to predict ozone concentrations for two different scenarios, emissions from a base case scenario and then projected emissions based on the technology and regulations that will be or are currently being implemented (in this case for 2000 and 2010). The model data show the expected impact of various



Regional Air Quality. Figure 13
Modeled ozone concentrations expressed as the sum of means over 35ppbV (SOMO35) for the year 2000 (*left*) and for the year 2010 (*right*) for the Air for Europe (CAFE) baseline scenario, which takes into account changes in air quality standards and reductions in emissions (IIASA; EEA [39])



© Crown copyright. All rights reserved Defra, Licence number 100018880 2007.

Regional Air Quality. Figure 14
Spatially disaggregated emissions of ammonia in the UK from 2008 (Murrells et al. [8])

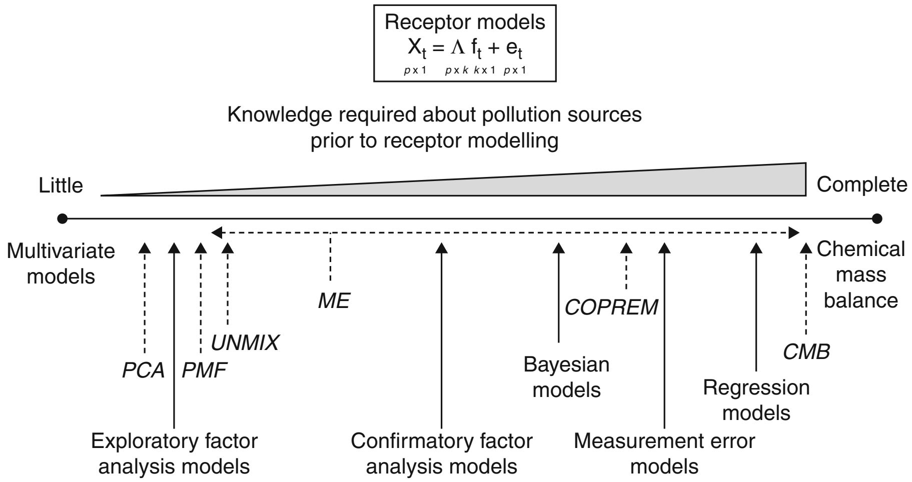
emission legislation measures, including legislation regarding combustion plants, the Euro standards for vehicles and non-road machinery, as well as International Panel on Climate Change (IPCC) and national legislation. As is visible in Fig. 13, the model predicts the largest impacts in ozone reduction will occur in the Mediterranean region [39].

The input data for many models are based on emission inventories. Emission inventories, like models, exist for different scales, i.e., local, regional, and global and resolutions, ranging from 1 km × 1 km to 1° × 1° (which is approximately 110 km × 110 km at the largest, depending on latitude). The species included in many emission inventories are greenhouse gases, nitrogen oxides, carbon monoxide, methane, non-methane volatile organic compounds, sulfur dioxide, ammonia, particulate matter, and recently black carbon and organic carbon [36]. Emission sources in inventories include anthropogenic emissions from fuel production, industrial and domestic combustion, transportation, waste disposal, industrial processes, solvent production and use, and agriculture, and biogenic emissions from vegetation and dust, and biomass burning emissions, which can be natural or anthropogenic [36]. Anthropogenic, natural, and biomass burning emissions are typically compiled in separate inventories. Examples of emission inventory data are given in Fig. 3 for carbon monoxide and Fig. 14 for ammonia [8].

Statistical Models

Statistical models are empirical or semiempirical models based on the statistical analysis of measured data. Instead of following the evolution of emissions in the atmosphere, statistical models work backward, from measured data collected at a given location, a receptor site, and work out the responsible emission sources. These models are also frequently referred to as receptor models. There are a number of different approaches that are used based on the understanding of source–receptor relationships including the chemical mass balance (CMB) model, which requires complete knowledge of sources and results in a quantitative output of source contributions and uncertainties, and factor analysis methods, such as Principal Component Analysis (PCA), which require no a priori source information. The different types of receptor models and the amount of knowledge required for model use is summarized in Fig. 15 [63, 64].

Chemical mass balance models are frequently used to support air quality decision-making for policy, such as was the case for CMB studies in Los Angeles, CA, and Las Vegas, NV which were used for the creation of State Implementation Plans to attain PM_{10} requirements [65]. In addition, CMB can be used to assess the success of implemented control technologies, and which sources are potentials for reduction controls. Such applications are possible in that CMB results yield quantitative source contributions to the measured metric, i.e., $PM_{2.5}$, organic carbon, or total mass. From such results the sources with the largest contributions


Receptor models
$\underset{p\times 1}{X_t} = \underset{p\times k}{\Lambda}\,\underset{k\times 1}{f_t} + \underset{p\times 1}{e_t}$
Knowledge required about pollution sources prior to receptor modelling
Little
Complete
Multivariate models
Chemical mass balance
ME
COPREM
UNMIX
PCA
PMF
CMB
Bayesian models
Regression models
Exploratory factor analysis models
Confirmatory factor analysis models
Measurement error models

Regional Air Quality. Figure 15

Approaches for estimating pollution source contributions using receptor models. Specific models are shown in italics and with dotted arrows (Schauer et al. [63]; Viana et al. [64])

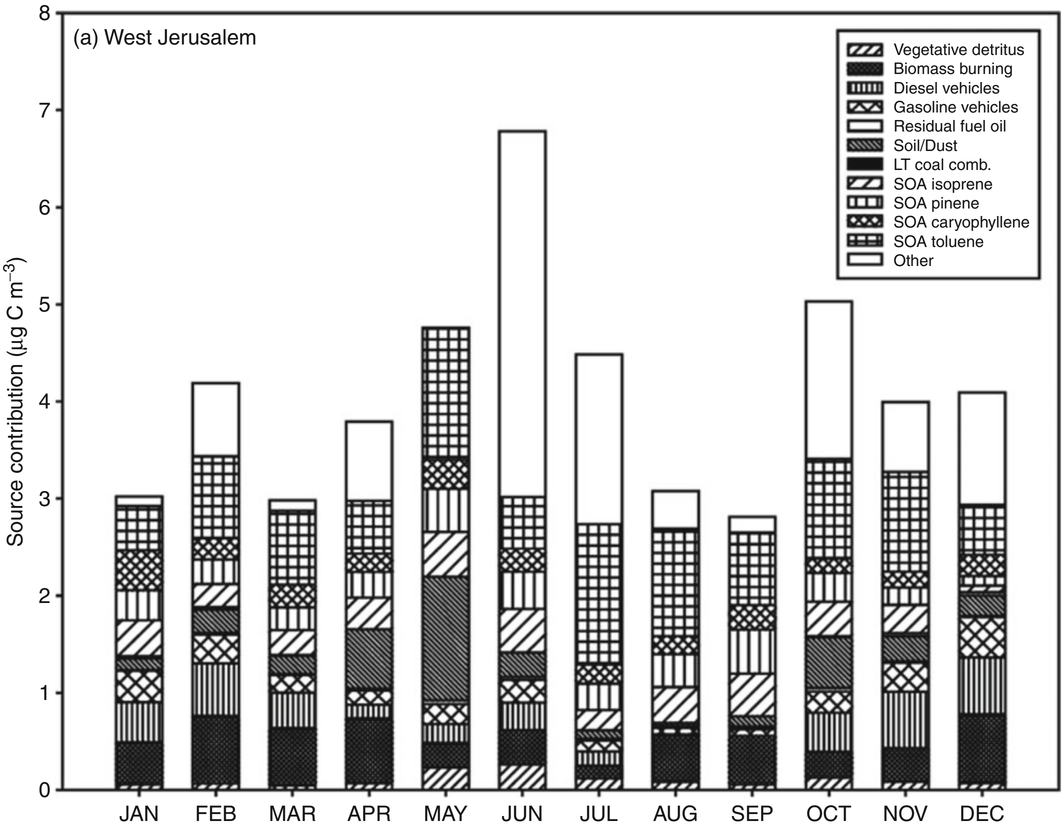
can be targeted for reduction, or changes in specific source contributions assessed. Recent reviews of CMB and other receptor models found that for the studies analyzed (the majority of which were from measurements in the USA and Europe), fossil fuel combustion is an important contributor to PM concentrations, with the primary contributions stemming from gasoline and diesel vehicle exhaust [64, 65]. Stationary sources, such as power stations, have only minimal source contributions when the facilities have effective pollution controls in place, but can be large contributors without such control technologies [65]. An example of the type of result produced by CMB modeling where fine-particulate matter organic carbon mass is apportioned is shown in Fig. 16 for West Jerusalem in 2007 [66].

In addition to receptor models, other types of statistical models are also frequently used in atmospheric chemistry, such as simple regression analysis, land use regression models, and neural network models. Land use regression models have been used in atmospheric chemistry to characterize air pollution exposure and health effects for people residing in urban areas, integrating traffic and geographic information into the models [67]. To learn about neural network models and their uses in the air pollution modeling field the reader is referred to Boznar and Mlakar, 2002 [68].

Current Air Quality Legislation and Metrics

Air quality guidelines are established to protect human health and the environment. The World Health Organization (WHO) has established air quality guidelines for use worldwide that are focused on protecting human health. Air quality guidelines in the USA and Europe are intended to protect human health, including sensitive populations (e.g., elderly, asthmatics), with secondary limits set for the protection of public welfare (e.g., visibility, animals, crops, buildings) [37]. The standards are set based on the extensive amount of research available concerning air pollution and health effects at the time of regulation.

The guidelines set by the WHO exist for the most common air pollutants: size-resolved particulate matter, ozone, nitrogen dioxide, and sulfur dioxide [1]. These were most recently updated in 2005. While the WHO has set guidelines for all countries to use as a basis for regulation if they choose to do so, most national air monitoring authorities have a list of "air quality pollutants" that are monitored and regulated. In the USA, these are the "criteria pollutants" that are regulated under the National Ambient Air Quality Standards (NAAQS) by the US Environmental Protection Agency (EPA) and include carbon monoxide, lead, nitrogen dioxide, particulate matter of two-size fraction, ozone,



Regional Air Quality. Figure 16
Source apportionment by chemical mass balance (CMB) modeling of monthly averaged ambient fine organic aerosol mass concentrations for West Jerusalem for 2007. LT Coal Comb = low-temperature coal combustion; SOA = secondary organic aerosol from the named precursor compound (e.g., isoprene) (von Schneidemesser et al. [66])

and sulfur dioxide, similar to those pollutants for which the WHO has established guidelines [37]. European Union air quality pollutants include particulate matter, sulfur dioxide, nitrogen dioxide, lead, carbon monoxide, benzene, ozone, arsenic, cadmium, nickel, and PAHs and are regulated under the air quality standards which were simplified in 2008 by combining all existing legislation on air quality under one directive known as the ambient air quality and Clean Air for Europe (CAFE) directive [69]. As is apparent here, all air quality pollutant regulation tends to contain a similar set of air pollutants with some countries/regulating bodies selecting additional species, such as trace metals, to be included. Regulatory standards and WHO guidelines are listed in Table 1.

Metrics also exist to quantify health or environmental effects of air pollutants, with different metrics established for human health and the environment. For example, AOT40 is the accumulated exposure over a threshold of 40 ppbV during daylight hours for a relevant period. SOMO35 is the sum of the daily maximum of running 8-h means over 35 ppbV. Both the AOT40 and SOMO35 are ozone metrics. Metrics such as AOT40 and SOMO35 are frequently considered to be more appropriate measures of the effects of ozone on crops and vegetation than those metrics designed to protect human health. A sample of environmental and health metrics are given in Table 2 [70].

Regional Context

There are a variety of spatial scales that both influence air quality and that air quality impact, such as the local, regional, and global. It is clear that these scales are not

Regional Air Quality. Table 1 Annual mean (unless noted otherwise) air quality guidelines as established by the World Health Organization (WHO), the United States Environmental Protection Agency (US EPA), and the European Union (Euro Directives)

	WHO	US EPA	Euro directive
Particulate matter (PM_{10})	20 μg m^{-3}	150 μg m^{-3} (24 h)	40 μg m^{-3}
Particulate matter ($PM_{2.5}$)	10 μg m^{-3}	15.0 μg m^{-3}	25 μg m^{-3}
Nitrogen dioxide	40 μg m^{-3}	53 ppb (~100 μg m^{-3})	40 μg m^{-3}
Sulfur dioxide	20 μg m^{-3} (24 h)	0.03 ppm (~80 μg m^{-3})	125 μg m^{-3}
Carbon monoxide		10 mg m^{-3} (8 h)	10 mg m^{-3} (8 h)
Lead		1.5 μg m^{-3} (quarterly average)	0.5 μg m^{-3}
Benzene			5 μg m^{-3}
Arsenic			6 ng m^{-3}
Cadmium			5 ng m^{-3}
Nickel			20 ng m^{-3}
Polycyclic aromatic hydrocarbons			1 ng m^{-3}

independent and definitions are loose. The local scale tends to the anthropogenic emission/human influence scale encompassing a city or a smaller area within a city, while regional scales are typically larger, encompassing an entire metropolitan region and the surrounding area or even greater. An example of a regional scale could be London and southeast England, or it could also be as large as all of Europe, depending on how the user defines it. When considering air quality on any regional scale, secondary pollutants from atmospheric processing and the transport of pollutants become much more important and primary emissions less so.

For example, as discussed earlier, ozone is an important secondary pollutant, which because of atmospheric processing time, has significant implications for regional air quality. Plume transport for urban areas into the surrounding region can have a significant influence on the air quality in the surrounding region [71]. In many cases, ozone concentrations have been observed to be higher in rural areas than urban areas [72, 73]. Furthermore, since regional air quality frequently includes rural areas, and thereby areas with fewer primary anthropogenic emissions, natural emissions can play a more significant role. For example, biogenic VOCs, such as isoprene, can contribute to ozone and other secondary pollutant formation, and wind erosion can add soil and/or dust to the existing aerosol loading. While these factors can and do also exist on a local scale, their relative importance on a local or regional scale varies depending on the location and size of the areas chosen.

Future Directions

Air Quality and Climate

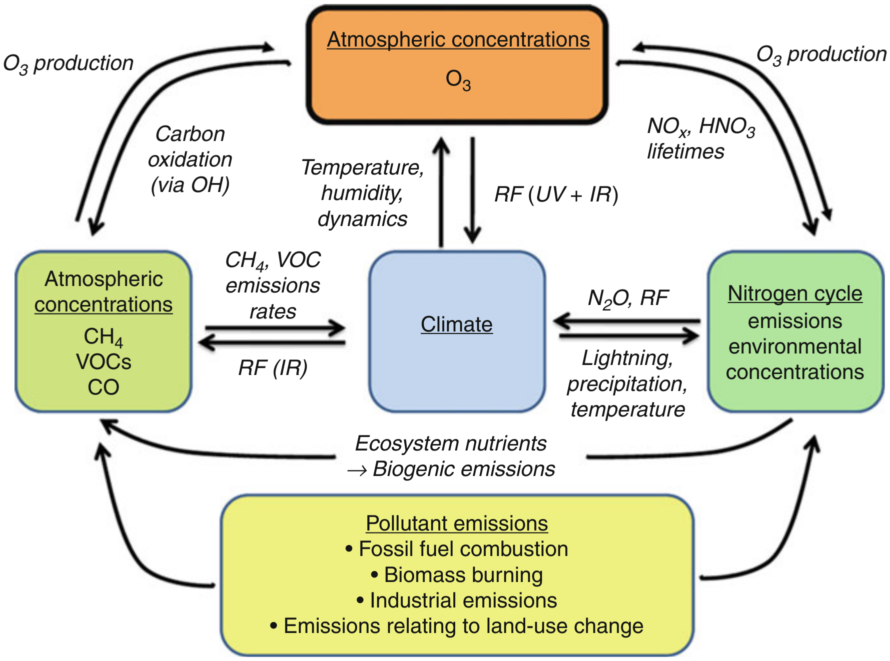
Air quality and climate change are important issues with complex interactions (see Figs. 1 and 17). Both are environmental issues that are and will continue to be addressed by environmental policies in the coming years. Much of the current focus is on trade-offs between different scenarios. While there are win-win situations where policy will improve both air quality and reduce greenhouse gas emissions, many measures may not result in such co-benefits. As stated in the IPCC Fourth Assessment Report "Future climate change may cause significant air quality degradation by changing the dispersion rate of pollutants, the chemical environment for ozone and aerosol generation and the strength of emissions from the biosphere, fires, and dust. The sign and magnitude of these effects are highly uncertain and will vary regionally" [74]. Ozone and aerosols are the two air pollutants of

Regional Air Quality. Table 2 A sample of ozone metrics of relevance to human health (Reproduced from AQEG 2009)

Metric	Relevance	Key influences on the values of this metric at urban locations
Annual average	Basic metric used to show long-term trends	Includes all of the hours in the year. Strongly influenced by the magnitude of local NO_x emissions and by topography through nocturnal depletion
Annual average of the daily maximum of the running 8-h mean	Used as "basic metric" for many of the health metrics.	Strongly influenced by the magnitude of local NO_x emissions
Annual average of the daily maximum of the running 8-h mean with a 70 $\mu g\ m^{-3}$ cutoff	Health impact, related to SOMO35	Influenced by the magnitude of local NO_x emissions and by photochemical episodes
Annual average of the daily maximum of the running 8-h mean with a 100 $\mu g\ m^{-3}$ cutoff	Health impact	Strongly influenced by photochemical episodes and to a lesser extent by the magnitude of local NO_x emissions
Maximum 1-h average (peak hour in the year)	Used as the basis for some epidemiological studies, although it has been suggested that the 8-h metric is more representative. Also an indicator of short-term peaks, but note low statistical power, since it is the value for one single hour	The metric most sensitive to the magnitude of regionally generated photochemical episodes and thus likely to show a response to reductions in relevant precursor emissions
Number of days with daily maximum of running 8-h mean exceeding 100 $\mu g\ m^{-3}$	Equates to the number of exceedences of the UK ozone standard (the Air Quality Strategy objective is no more than 10 exceedences per year)	Strongly influenced by photochemical episodes and to a lesser extent by the magnitude of local NO_x emissions
Number of days with daily maximum of running 8-h mean exceeding 120 $\mu g\ m^{-3}$	Equates to the number of exceedences of the EU Target Value (no more than 25 days, averaged over 3 years) and Long-Term Objective (no exceedences) from the third Daughter Directive	Strongly influenced by photochemical episodes and to a lesser extent by the magnitude of local NO_x emissions
SOMO35 (sum of means over 35 ppb)	Used as a metric by IIASA, for Clean Air for Europe (CAFE) and NECD revision, related to annual average of the daily maximum of the running 8-h mean with a 70 $\mu g\ m^{-3}$ cutoff	Influenced by the magnitude of local NO_x emissions and by photochemical episodes

greatest concern for public health [75], while being key players in climate change and feedbacks [36, 76].

Air pollution and greenhouse gases often have common sources of emission. In addition to the common sources, there are some gases which are both air pollutants and act as greenhouse gases, such as ozone [77]. To best address both issues, a significant effort will need to be made to understand the impact of a changing climate on air quality and vice versa. Current work with global and regional models that couple chemical transport and general circulation simulations have found that climate change alone could increase summertime ozone concentrations and the severity of summer smogs over the coming decades, with the largest effects in polluted, urban areas [75–77]. The potential impact of climate change on PM is much less certain. Important factors in changes in PM owing to climate change include precipitation frequency and mixing depth, as


Atmospheric concentrations
O_3
O_3 production
O_3 production
Carbon oxidation (via OH)
Temperature, humidity, dynamics
RF (UV + IR)
NO_x, HNO_3 lifetimes
Atmospheric concentrations
CH_4
VOCs
CO
CH_4, VOC emissions rates
RF (IR)
Climate
N_2O, RF
Lightning, precipitation, temperature
Nitrogen cycle
emissions
environmental concentrations
Ecosystem nutrients
→ Biogenic emissions
Pollutant emissions
• Fossil fuel combustion
• Biomass burning
• Industrial emissions
• Emissions relating to land-use change

Regional Air Quality. Figure 17

Schematic representation of the multiple interactions between tropospheric chemical processes, biogeochemical cycles, and the climate system. *RF* radiative forcing, *UV* ultraviolet radiation, *IR* infrared radiation (IPCC, 2007; EPA [76])

well as wildfire frequencies [75, 76]. In addition to the uncertainty of the impact of climate change on PM emissions, the uncertainty of the reverse, the impact of PM on climate change, is also large. Part of this uncertainty arises from the differences in aerosol types and their climate properties. Sulfates and nitrates have an overall cooling (negative radiative forcing) effect, whereas black carbon has an overall warming effect [36]. Therefore, while the reduction of PM improves air quality, it may result in further atmospheric warming in terms of climate change. In addition to direct forcing, aerosols can also act as cloud condensation nuclei (CCN) and exert an indirect effect on climate. The abundance of CCN affects cloud formation and cloud lifetime, which in turn affects the scattering and absorption of radiation and therefore climate. Preliminary model results show a range of increases and decreases in PM concentrations for different regions in response to climate change [75, 76].

While controls of CO_2 emissions should remain a priority for mitigation of global climate change, reductions in emissions of air pollutants could have a faster impact on slowing warming trends in the short term (10–30 years). Raes and Seinfeld [78] have pointed out that we maybe on a "bumpy road" to recovery with larger short-term increases in temperature, owing partly to the fact that significant amounts of long-lived GHGs are already in the atmosphere but also because reductions in atmospheric aerosols in the next decade or so will add to the positive radiative forcing of the Earth's climate.

Mitigation of short-lived agents with high radiative forcing may reduce the possibility of catastrophic climate change in particularly sensitive regions such as the Arctic. However, mitigation decisions must be based on sound science and would benefit from a sector-based approach accounting for benefits for health, ecosystems, climate, and mitigation costs. While the ideal goal is to achieve win-win solutions for air pollution and climate, in some cases health improvements via reductions in pollutant emissions may greatly outweigh possible negative effects on climate. This is particularly true of the developing world where improving health and ensuring food security currently have immediate priority over the benefits of slowing the rate of climate change. Nevertheless, the development of emission reduction strategies or cleaner energy alternatives for both CO_2 and short-lived climate forcers (SLCFs) are needed. There are clearly many challenges in the area of air quality and climate and a need for sound underpinning science across the spatial and temporal scales that air quality and climate act on.

Bibliography

Primary Literature

1. WHO (2005) WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide. World Health Organization Regional Office for Europe, Copenhagen
2. Jacobson MZ (2002) Atmospheric pollution history, science, and regulation. Cambridge University Press, Cambridge
3. Fenger J (2009) Air pollution in the last 50 years – from local to global. Atmos Environ 43:13–22
4. Seinfeld JH (2004) Air pollution: a half century of progress. Environ Energy Eng 50(6):1096–1108
5. Seinfeld JH, Pandis SN (2006) Atmospheric chemistry and physics: from air pollution to climate change, 2nd edn. Wiley, New York
6. Bradley KS, Stedman DH, Bishop GA (1999) A global inventory of carbon monoxide emissions from motor vehicles. Chemosphere Glob Change Sci 1:65–72
7. von Schneidemesser E, Monks PS, Duelmer PC (2010) Global comparison of VOC and CO observations in urban areas. Atmos Environ 44(37):4772–4816
8. Murrells T, Passant N, Thistlethwaite G, Wagner A, Li Y, Bush T, Norris J, Walker C, Stewart R, Tsagatakis I, Whiting R, Conolly C, Okamura S, Peirce M, Sneddon S, Webb J, Thomas J, MacCarthy J, Choudrie S, Brophy N (2010) UK Emissions of Air Pollutants 1970 to 2008. AEA
9. Clerbaux C, Edwards DP, Deeter M, Emmons L, Lamarque J-F, Tie XX, Massie ST, Gille J (2008) Carbon monoxide pollution from cities and urban areas observed by the Terra/MOPITT mission. Geophys Res Lett 35:L03817. doi:10.1029/2007GL032300
10. EPA U (2010) Our Nation's air, status and trends through 2008. United States Environmental Protection Agency
11. Finlayson-Pitts BJ, Pitts JN Jr (2000) Chemistry of the upper and lower atmosphere: theory, experiments, and applications. Academic, San Diego/London
12. Gurjar BR, Butler TM, Lawrence MG, Lelieveld J (2008) Evaluation of emissions and air quality in megacities. Atmos Environ 42:1593–1606
13. Guttikunda SK, Tang Y, Charmichael GR, Kurata G, Pan L, Streets DG, Woo J-H, Thongboonchoo N, Fried A (2005) Impacts of Asian megacity emissions on regional air quality during spring 2001. J Geophys Res 110:D20301
14. Leighton PA (1961) The photochemistry of air pollution. Academic, New York
15. Soltic P, Weilenmann M (2003) NO2/NO emissions of gasoline passenger cars and light-duty trucks with Euro-2 emission standard. Atmos Environ 37:5207–5216
16. Carslaw DC (2005) Evidence of an increasing NO2/NOx emissions ratio from road traffic emissions. Atmos Environ 39:4793–4802
17. Stephens S, Madronich S, Wu F, Olson JB, Ramos R, Retama A, Munoz R (2008) Weekly patterns of Mexico City's surface concentrations of CO, NOx, PM10 and O3 during 1986–2007. Atmos Chem Phys 8:5313–5325
18. Godish T (2004) Air quality, 4th edn. Lewis, Boca Raton
19. Jacob DJ (1999) Introduction to atmospheric chemistry. Princeton University Press, Princeton
20. Watson JG, Chow JC, Fujita EM (2001) Review of volatile organic compound source apportionment by chemical mass balance. Atmos Environ 35:1567–1584
21. Baker AK, Beyersdorf AJ, Doezema LA, Katzenstein A, Meinardi S, Simpson IJ, Blake DR, Rowland FS (2008) Measurements of nonmethane hydrocarbons in 28 United States cities. Atmos Environ 42:170–182
22. Borbon A, Locoge N, Veillerot M, Galloo JC, Guillermo R (2002) Characterisation of NMHCs in a French urban atmosphere: overview of the main sources. Sci Total Environ 292:177–191
23. Bower J, Broughton G, Connolly C, Cook A, Grice S, Kent A, Loader A, Stedman J, Targa J, Telling S, Tsagatakis I, Vincent K, Willis P, F-w Y, Yardley R (2009) Air pollution in the UK: 2008. AEA, Didcot
24. Dollard GJ, Dore CJ, Jenkin ME (2001) Ambient concentrations of 1, 3-butadiene in the UK. Chem-Biol Interact 135–136:177–206
25. Barletta B, Meinardi S, Roland FS, Chan C-Y, Wang X, Zou S, Chan LY, Blake DR (2005) Volatile organic compounds in 43 Chinese cities. Atmos Environ 39:5979–5990
26. Fischer H, Nikitas C, Parchatka U, Zenker T, Harris GW, Matuska P, Schmitt R, Mihelcic D, Muesgen P, Paetz HW, Schultz M, Thomas VA (1998) Trace gas measurements during the oxidizing capacity of the tropospheric atmosphere campaign 1993 at Izana. J Geophys Res 103(D11):13505–513518
27. Vingarzan R (2004) A review of surface ozone background levels and trends. Atmos Environ 38:3431–3442
28. Oltmans SJ, Lefohn AS, Harris JM, Shadwick DS (2008) Background ozone levels of air entering the west coast of the US and assessment of longer-term changes. Atmos Environ 42(24):6020–6038. doi:10.1016/j.atmosenv.2008.03.034
29. Chan CK, Yao X (2008) Air pollution in mega cities in China. Atmos Environ 42(1):1–42. doi:10.1016/j.atmosenv.2007.09.003
30. de Leeuw F, Bogman F (2001) Air pollution by ozone in Europe in summer 2001. European Environment Agency, Copenhagen
31. de Leeuw FAAM, de Paus TA (2000) Exceedance of EC Ozone threshold values in Europe in 1997. Water Air Soil Pollut 128:255–281
32. Fowler D, Coyle M, Anderson R, Ashmore MR, Bower JS, Burgess RA, Cape JN, Cox RA, Derwent RG, Dollard GJ, Grennfelt P, Harrison RM, Hewitt CN, Hov O, Jenkin ME, Lee DS,

Maynard RL, Penkett SA, Smith RI, Stedman JR, Weston KJ, Williams ML, Woods PJ (1997) Ozone in the United Kingdom. Air and Environment Quality Division, Department of the Environment, Transport and the Regions, London, UK

33. Geddes JA, Murphy JG, Wang DK (2009) Long term changes in nitrogen oxides and volatile organic compounds in Toronto and the challenges facing local ozone control. Atmos Environ 43:3407–3415
34. Sather ME, Cavender K (2007) Trends analysis of ambient 8 hour ozone and precursor monitoring data in the south central US. J Environ Monit 9(2):143–150. doi:10.1039/b617603h
35. Carvalho A, Monteiro A, Ribeiro I, Tchepel O, Miranda AI, Borrego C, Saavedra S, Souto JA, Casares JJ (2010) High ozone levels in the northeast of Portugal: analysis and characterization. Atmos Environ 44:1020–1031
36. Monks PS, Granier C, Fuzzi S, Stohl A, Williams ML, Akimoto H, Amann M, Baklanov A, Baltensperger U, Bey I, Blake N, Blake RS, Carslaw K, Cooper OR, Dentener F, Fowler D, Fragkou E, Frost GJ, Generoso S, Ginoux P, Grewe V, Guenther A, Hansson HC, Henne S, Hjorth J, Hofzumahaus A, Huntrieser H, Isaksen ISA, Jenkin ME, Kaiser J, Kanakidou M, Klimont Z, Kulmala M, Laj P, Lawrence MG, Lee JD, Liousse C, Maione M, McFiggans G, Metzger A, Mieville A, Moussiopoulos N, Orlando JJ, O'Dowd CD, Palmer PI, Parrish DD, Petzold A, Platt U, Poschl U, Prevot ASH, Reeves CE, Reimann S, Rudich Y, Sellegri K, Steinbrecher R, Simpson D, ten Brink H, Theloke J, van der Werf GR, Vautard R, Vestreng V, Vlachokostas C, von Glasow R (2009) Atmospheric composition change – global and regional air quality. Atmos Environ 43(33):5268–5350. doi:10.1016/j.atmosenv.2009.08.021
37. EPA U (2010) United States Environmental Protection Agency National Ambient Air Quality Standards. http://epa.gov/air/criteria.html
38. Pope CAI, Dockery DW (2006) Health effects of fine particulate air pollution: lines that connect. J Air Waste Manage Assoc 56:709–742
39. EEA (2007) Air pollution in Europe 1990–2004. EEA Report No2/2007. European Environment Agency, Copenhagen
40. Tsigaridis K, Krol M, Dentener F, Balkanski Y, Lathiere J, Metzger S, Hauglustaine D, Kanakidou M (2006) Change in global aerosol composition since preindustrial times. Atmos Chem Phys 6:5143–5162
41. Colvile R, Hutchinson E, Mindell J, Warren R (2001) The transport sector as a source of air pollution. Atmos Environ 35:1537–1565
42. Hewitt CN, Jackson AV (eds) (2009) Atmospheric science for environmental scientists. Blackwell, Chichester
43. Williams PI, Baltensperger U (2009) Particulate matter in the atmosphere. In: Hewitt CN, Jackson AV (eds) Atmospheric science for environmental scientists. Blackwell, Chichester
44. Poeschl U (2005) Atmospheric aerosols: composition, transformation, climate and health effects. Angew Chem Int Ed 44:7520–7540
45. Stone EA, Hedman CJ, Zhou JB, Mieritz M, Schauer JJ (2010) Insights into the nature of secondary organic aerosol in Mexico City during the MILAGRO experiment 2006. Atmos Environ 44(3):312–319. doi:10.1016/j.atmosenv.2009.10.036
46. Stone EA, Zhou JB, Snyder DC, Rutter AP, Mieritz M, Schauer JJ (2009) A comparison of summertime secondary organic aerosol source contributions at contrasting urban locations. Environ Sci Technol 43(10):3448–3454. doi:10.1021/es8025209
47. von Schneidemesser E, Zhou JB, Stone EA, Schauer JJ, Qasrawi R, Abdeen Z, Shpund J, Vanger A, Sharf G, Moise T, Brenner S, Nassar K, Saleh R, Al-Mahasneh QM, Sarnat JA (2010) Seasonal and spatial trends in the sources of fine particulate organic carbon in Israel, Jordan, and Palestine. Atm Environ 44:3669–3678
48. Fuzzi S, Andreae MO, Huebert BJ, Kulmala M, Bond TC, Boy M, Doherty SJ, Guenther A, Kanakidou M, Kawamura K, Kerminen VM, Lohmann U, Russell LM, Poschl U (2006) Critical assessment of the current state of scientific knowledge, terminology, and research needs concerning the role of organic aerosols in the atmosphere, climate, and global change. Atmos Chem Phys 6:2017–2038
49. Stone E, Schauer J, Quraishi TA, Mahmood A (2010) Chemical characterization and source apportionment of fine and coarse particulate matter in Lahore, Pakistan. Atmos Environ 44:1062–1070
50. Toscano G, Gambaro A, Moret I, Capodaglio G, Turetta C, Cescon P (2005) Trace metals in aerosol at Terra Nova Bay, Antarctica. J Environ Monit 7:1275–1280
51. von Schneidemesser E, Stone EA, Quraishi TA, Shafer MM, Schauer JJ (2010) Toxic metals in the atmosphere in Lahore, Pakistan. Sci Total Environ 408(7):1640–1648. doi:10.1016/j.scitotenv.2009.12.022
52. Sillanpaa M, Frey A, Hillamo R, Pennanen AS, Salonen RO (2005) Organic, elemental and inorganic carbon in particulate matter of six urban environments in Europe. Atmos Chem Phys 5:2869–2879
53. European Monitoring and Evaluation Programme (2010) EMEP Convention on long-range transboundary air pollution. http://www.emep.int/. Accessed July 2010
54. EMEP (2009) Progress in activities in 2009 and future work: measurements and modelling (acidifications, eutrophication, photo-oxidants, heavy metals, particulate matter and persistant organic pollutants). Geneva
55. Fenger J (2009) Urban air pollution. In: Hewitt CN, Jackson AV (eds) Atmospheric science for environmental scientists. Blackwell, Chichester
56. Sillman S (1999) The relation between ozone, NOx and hydrocarbons in urban and polluted rural environments. Atmos Environ 33:1821–1845
57. Brunekreef B, Holgate ST (2002) Air pollution and health. Lancet 360:1233–1242
58. Dockery DW, Pope CA III, Xu X, Spengler JD, Ware JH, Fay ME Jr, Ferris BG, Speizer FE (1993) An association between air pollution and mortality in six US Cities. N Eng J Med 329(24):1753–1759
59. Pope CAI, Thun MJ, Namboodiri MM, Dockery DW, Evans JS, Speizer FE, Heath JCW (1995) Particulate air pollution as a predictor of mortality in a prospective study of U.S. adults. Am J Respir Crit Care Med 151:669–674
60. COMEAP (2009) Long-term exposure to air pollution: effect on mortality. Health Protection Agency for the Committee on

the Medical Effects of Air Pollutants, Chilton, Didcot, Oxfordshire, UK

61. Pope CAI, Ezzati M, Dockery DW (2009) Fine-particulate air pollution and life expectancy in the United States. N Engl J Med 360(4):376–386
62. Harrison RM (ed) (1999) Understanding our environment. The Royal Society of Chemistry, Cambridge
63. Schauer JJ, Lough GC, Shafer MM, Christensen WF, Arndt MF, DeMinter JT, Park J-S (2006) Characterization of metals emitted from motor vehicles. HEI Research Report, vol 133. Health Effects Institute, Boston
64. Viana M, Kuhlbusch TAJ, Querol X, Alastuey A, Harrison RM, Hopke PK, Winiwarter W, Vallius A, Szidat S, Prevot ASH, Hueglin C, Bloemen H, Wahlin P, Vecchi R, Miranda AI, Kasper-Giebl A, Maenhaut W, Hitzenberger R (2008) Source apportionment of particulate matter in Europe: a review of methods and results. J Aerosol Sci 39(10):827–849. doi:10.1016/j.jaerosci.2008.05.007
65. Chow JC, Watson JG (2002) Review of PM2.5 and PM10 apportionment for fossil fuel combustion and other sources by the chemical mass balance receptor model. Energy Fuels 16(2):222–260. doi:10.1021/ef0101715
66. von Schneidemesser E, Zhou JB, Stone EA, Schauer JJ, Shpund J, Brenner S, Qasrawi R, Abdeen Z, Sarnat JA (2010) Spatial variability of carbonaceous aerosol concentrations in East and West Jerusalem. Environ Sci Technol 44:1911–1917
67. Ryan PH, LeMasters GK (2007) A review of land-use regression models for characterizing intraurban air pollution exposure. Inhal Toxicol 19(Suppl 1):127–133
68. Boznar MZ, Mlakar P (2002) Use of neural networks in the field of air pollution modelling. In: Borrego C, Shayes G (eds) Air pollution modelling and its application XV. Kluwer, Secaucus
69. EC (2010) European commission environment air quality standards. http://ec.europa.eu/environment/air/quality/standards.htm
70. AQEG (2009) Ozone in the United Kingdom. Department for the Environment, Food and Rural Affairs, London
71. Penkett S, Clemitshaw K, Savage N, Burgess R, Cardenas L, Carpenter L, McFadyen G, Cape J (1999) Studies of oxidant production at the weybourne atmospheric observatory in summer and winter conditions. J Atmos Chem 33:111–128
72. Jenkin ME (2008) Trends in ozone concentration distributions in the UK since 1990: local, regional and global influences. Atmos Environ 42:5434–5445
73. EEA (2001) Air pollution by ozone in Europe in summer 2001. European Environment Agency, Copenhagen
74. Denman KL, Brasseur G, Chidthaisong A, Ciais P, Cox PM, Dickinson RE, Hauglustaine D, Heinze C, Holland E, Jacob D, Lohmann U, Ramachandran S, da Silva Dias PL, Wofsy SC, Zhang X (2007) Couplings between changes in the climate system and biogeochemistry. climate change 2007: the physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge/New York
75. Jacob DJ, Winner DA (2009) Effect of climate change on air quality. Atmos Environ 43:51–63
76. EPA U (2009) Assessment of the impacts of global change on regional U.S. air quality: a synthesis of climate change impacts on ground-level ozone. National Center for Environmental Assessment, Washington DC
77. AQEG (2010) Air pollution: action in a changing climate. Department for Environment Food and Rural Affairs, London
78. Raes F, Seinfeld JH (2009) New directions: climate change and air pollution abatement: a bumpy road. Atmos Environ 43:5132–5133
79. Whitby KT, Cantrell BK (1976) Atmospheric Aerosols: Characteristics and Measurement. International Conference on Environmental Sensing and Assessment (ICESA), Institute of Electrical and Electronic Engineers (IEEE). Las Vegas, NV. September 14–19, 1975

Books and Reviews

Brasseur GP, Orlando JJ, Tyndall GS (1999) Atmospheric chemistry and global change. Oxford University Press, Oxford

Brimblecombe P, Hara H, Houle D, Novak M (2007) Acid rain-deposition to recovery. Springer, Dordrecht

IPCC (2007) Climate change 2007 – The physical science basis; Contribution of working group I to the fourth assessment report of the IPCC, Cambridge University Press, Cambridge

Finlayson-Pitts BJ, Pitts JN (2000) Chemistry of the upper and lower atmosphere. Academic, San Diego/London

Hewitt CN, Jackson AV (2003) Handbook of atmospheric science: principles and applications. Blackwell, Oxford

Holloway AM, Wayne RP (2010) Atmospheric chemistry. The Royal Society of Chemistry, Cambridge

Houghton JT (2002) The physics of atmospheres, 3rd edn. Cambridge University Press, Cambridge

Jacobson MZ (2005) Fundamentals of atmospheric modeling, 2nd edn. Cambridge University Press, New York

Seinfeld JH, Pandis SN (2006) Atmospheric chemistry and physics: from air pollution to climate change, 2nd edn. Wiley, New York

Wallace JM, Hobbs PV (2006) Atmospheric science: an introductory survey, 2nd edn. Academic, London/San Diego

Regional Climate Models

L. Ruby Leung
Atmospheric Sciences and Global Change Division,
Pacific Northwest National Laboratory, Richland, WA,
USA

Article Outline

Glossary
Definition of the Subject and Its Importance

Introduction
Modeling Approach
Evaluating Regional Climate Models
Application of Regional Climate Models
Future Directions
Acknowledgments
Bibliography

Glossary

Downscaling Development of climate information at local or regional scale from coarse resolution data or model outputs; both statistical and dynamical methods can be used.

GCM Global climate model, a climate model based on the general circulation of the atmosphere, often coupled with models of ocean circulation and sea ice.

Mesoscale In the atmosphere, mesoscale generally refers to horizontal scales that lie between the scale height of the atmosphere (about 10 km) and the Rossby radius of deformation (tens to hundreds of kilometers).

Nudging Method to reduce the differences between the simulated and observed or imposed states by applying corrections, usually in the form of tendencies to the prognostic equations, based on the differences.

RCM Regional climate model (also called nested regional climate model), a climate model applied over a limited area with boundary conditions provided by global models or analyses.

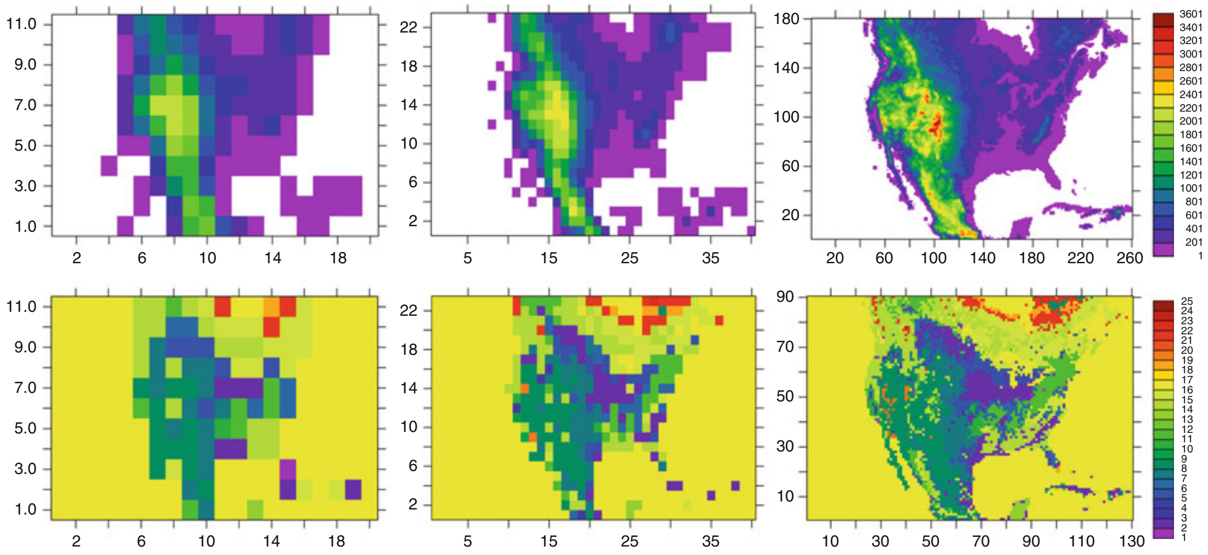
Definition of the Subject and Its Importance

Regional climate models are numerical models that simulate the climate of geographic regions typically covering a few thousand square kilometers to a continent. Most regional climate models include models that describe the atmosphere and the underlying land surface, but a few also include models of ocean and sea ice and atmospheric aerosols and chemistry. Given the atmospheric state at the lateral boundaries, regional climate models simulate regional climate in the context of the evolving global climate. Because regional domains cover only a fraction of the globe, it is computationally more feasible to apply regional climate models at higher grid resolution compared to global climate models to better resolve atmospheric and terrestrial processes and how they respond to regional forcings such as topography and land cover/land use. While global climate models are generally applied at grid resolution of a few hundred kilometers, regional climate models have been more commonly applied at grid resolution of a few tens of kilometers. Therefore, a common application of regional climate models is the dynamical downscaling of global climate simulations to provide regional climate information related to climate change projections or climate predictions. As such, regional climate models have served an important function of providing regional climate scenarios needed to assess a wide range of societal relevant climate impacts such as climate change effects on water resources and ecosystems. Regional climate models are also used to study regional climate processes, particularly those that are related to the water cycle that is inherently multi-scale; so explicitly representing finer scale processes is important to simulate its variations at multiple time and space scales.

Introduction

Regional climate models were first developed in the late 1980s to provide a means to simulate climate features that were not well captured by global climate models (GCMs) because of their coarse spatial resolution. Figure 1 shows the representation of surface elevation and land cover/land use in climate models of different horizontal resolutions. At 400 km resolution, which was typical for GCMs in the early 1990s, climate models can only resolve very crude topographic variations and land surface heterogeneities to simulate their effects on large-scale and mesoscale circulation. At 50 km resolution, which is a common resolution used in regional climate models even today, models can begin to realistically capture topographic and land cover features important for regional climate.

The first regional climate model (RCM) was developed and applied to the western USA where regional climate is significantly influenced by the complex terrain not well resolved by GCMs [13,17]. The RCM was adapted from a mesoscale or limited-area atmospheric model that was designed for weather forecasting or short-term simulation of a few days. The model was enhanced for climate simulation by improving the physics representations for processes such as radiative transfer and biosphere-atmosphere exchange at the land surface that governs the energy and water budgets



Regional Climate Models. Figure 1

Surface elevation (in meters) (*top row*) and land cover/land use (type) (*bottom row*) represented at 400 km (*left*), 200 km (*middle*), and 50 km (*right*) horizontal resolution in climate models. The land cover/land use types are: *1* urban, *2* dryland crop, *3* irrigated crop, *4* mixed crop, *5* crop/grass, *6* crop/woodland, *7* grass, *8* shrub, *9* mixed shrub/grass, *10* savanna, *11* deciduous broadleaf, *12* deciduous needleleaf, *13* evergreen broadleaf, *14* evergreen needleleaf, *15* mixed forest, *16* water bodies, *17* herbaceous wetland, *18* wooded wetland, *19* barren/sparsely vegetated, *20* herbaceous tundra, *21* wooded tundra, *22* mixed tundra, *23* bare ground tundra, *24* snow/ice, *25* playa. The x- and y-axes show the number of grid points in the domain at the three spatial resolutions

of the climate system. This was achieved by adopting the physics parameterizations used in a GCM. The RCM was driven at the lateral boundaries by atmospheric analysis [17] that provides an observationally constrained and dynamically balanced atmospheric state and global climate simulations [13].

Giorgi and Bates [17] showed, for the first time, that limited-area models could be used to produce long-term (more than a month) continuous simulations, as opposed to prior applications that use limited-area models to simulate weather for just a few days. By comparing the regional simulations with observations and the GCM simulations, it was demonstrated that a mesoscale weather model, with appropriate modifications, could be used for regional climate simulations. Following these pioneering studies, Giorgi et al. [19] further enhanced their RCM by updating the physics parameterizations with newer options available from the GCM, and explored model sensitivity to physics parameterizations and methods of assimilating the lateral boundary conditions. Giorgi and Mearns [16] showed that errors (e.g., measured by the deviation of the model solution from the driving large-scale fields) in limited-area models grow initially during model spin up, but reach an asymptotic value after a few days. At this stage, the climate simulated by the models is defined by the large-scale driving conditions and the model internal physics and dynamics, as well as the regional forcings within the model domain.

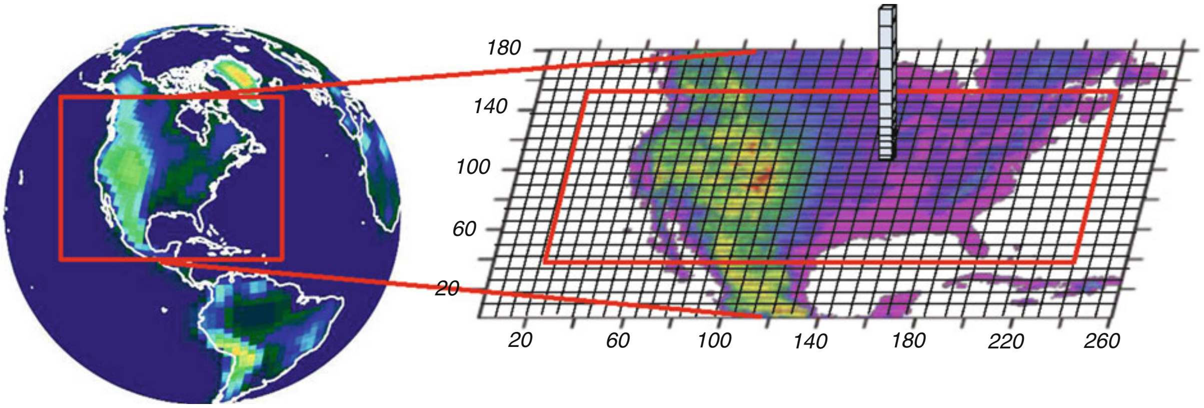
Subsequent to the early studies by Giorgi and his colleagues, more regional climate models have been developed following a similar approach and development path. These models have been applied to many regions around the world to assess their simulation skill under different climate regimes such as the monsoon, arid and semiarid deserts, mid-latitude regimes influenced by synoptic systems, and the high latitudes where cryospheric processes are important. As regional climate models became more widely used, questions have been raised about the validity and usefulness of the approach that prompted a series of studies to vigorously assess the various assumptions, and proposed practical or more mathematically well-posed solutions to regional climate modeling (section "Modeling Approach"). Different datasets and approaches have been used to evaluate RCMs, and large model intercomparison projects have been organized to evaluate and intercompare simulations produced by different RCMs (section "Evaluating Regional Climate Models"). At the same time, many studies have applied RCMs to simulate regional climate change that provided insights on climate change impacts. Regional climate models have also been used to study regional climate processes such as the role of land-atmosphere feedbacks on droughts and monsoon precipitation, effects of aerosols and land use on regional climate and the hydrological cycle, and processes leading to extreme climate events. The following sections provide a synopsis of these topics, and discuss the future directions in regional climate modeling. Examples of RCM applications are given in section "Application of Regional Climate Models."

Modeling Approach

How Do Regional Climate Models Work

Regional climate models are numerical models that simulate the climate of a specific region. Although some regional climate models, or regional earth system models, are beginning to include models of ocean, sea ice, and atmospheric aerosol and chemistry coupled to the atmosphere and land components, this review focuses mainly on regional climate models that traditionally include only atmosphere and land components with prescribed sea surface temperature and sea ice.

Similar to global atmospheric models, regional climate models numerically and simultaneously solve the equations of the conservation of energy, momentum, and water vapor that govern the atmospheric state. These equations are based on the Navier-Stokes equations for fluid flow (conservation of momentum) with approximations that apply to the atmosphere, the thermodynamic energy equation (conservation of energy), the continuity equation (conservation of mass), and the equation of state (ideal gas law). These partial differential equations are cast in various forms for different conservative properties and integrated forward in time using dynamical solvers. The solvers are applied to three-dimensional computational domains that are divided horizontally with grid spacing of a few to tens of kilometers and vertically into tens of vertical layers with a model top near 10–50 hPa. In regional climate models, solving these equations on limited-area domains require lateral boundary conditions, which can be derived from global climate simulations or global analyses to describe the large-scale atmospheric states. This method of simulating regional climate using limited-area models with prescribed lateral boundary conditions is called nesting (Fig. 2), so regional climate models are also called nested regional climate models to distinguish them



Regional Climate Models. Figure 2
A schematic showing the nesting of a regional climate model within a global climate model. The *right hand* figure shows the regional domain over North America with the horizontal grid (*black lines*), boundary of the buffer zone (*red box*), and a vertical column indicating the atmospheric layers represented by the model

from other dynamical frameworks such as global variable resolution or global stretched-grid models that simulate regional climate for specific regions through regional refinement within the global domain.

The most commonly used lateral boundary treatment in nested regional climate models involves the relaxation of the interior flow in the vicinity of the boundary, called the lateral boundary buffer zone, to the prescribed flow [8]. In most models, the same treatment is also applied to the thermodynamics variables. When applied to RCMs, increasing the width of the lateral boundary buffer zone allows stronger control of the lateral boundary conditions to keep the simulated large scales closer to the global simulations or analyses that provide the lateral boundary conditions. Some RCMs have the capability to use nesting to further zoom into smaller regions with increasing grid resolutions. As computational resources increased over time, more RCMs are now formulated using non-hydrostatic dynamics, as the mean vertical motion of the air column within a model grid cell can no longer be assumed negligible at higher grid resolution. In contrast, most GCMs use hydrostatic solvers because the hydrostatic assumption is valid in coarser grids.

Besides numerically solving the momentum, thermodynamics, and continuity equations, climate models, global or regional, include parameterizations of physical processes such as radiative transfer, convection, cloud microphysics, land surface and biosphere-atmosphere exchange, and boundary layer turbulence. These parameterizations calculate the diabatic heating, moistening, and momentum changes due to the various processes. The resulting tendencies or rates of change are included as sources and sinks in the equations of energy, momentum, and water vapor to drive the atmospheric circulation.

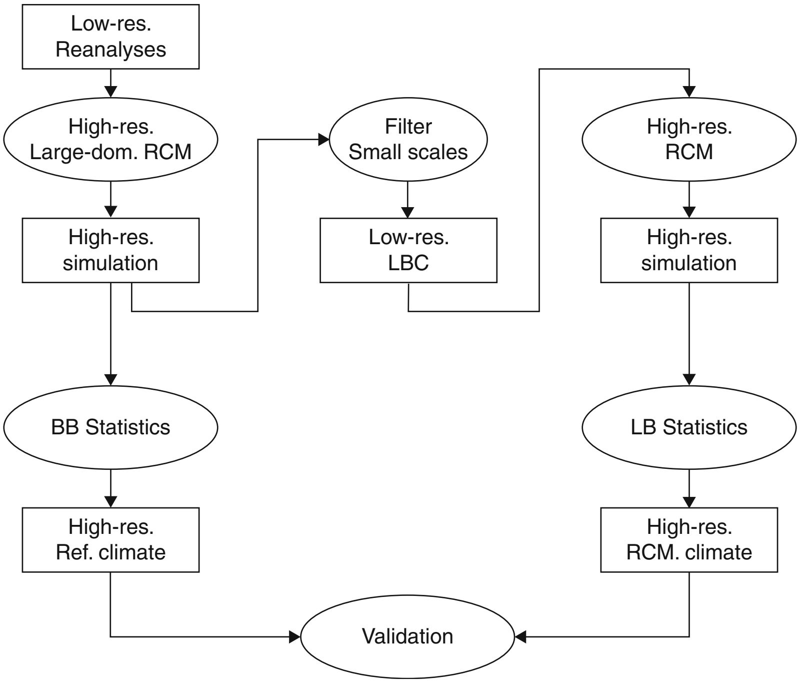
Traditionally, GCMs use more sophisticated parameterizations of slow physical processes such as radiation and land surface for more accurate simulations of the global energy budgets, while limited-area models that are developed mainly for weather forecasting and short-term simulations emphasize detailed parameterizations of fast physical processes such as cloud microphysics and turbulence transfer. To simulate regional climate, both fast and slow physical processes are important because of the short spatial scale and long time scale of interest. Therefore many RCMs have adapted parameterizations of slow processes from GCMs, while maintaining the suite of the relatively detailed parameterizations of fast processes used in weather forecasting. Sharing of physics parameterizations between the global and regional models is considered desirable to reduce inconsistency between the simulated and driving large-scale conditions (see section “Modeling Issues” for a discussion of potential issues caused by mismatch of GCM and RCM solutions) and facilitate interpretation of differences simulated by the RCMs and GCMs. Since the first RCM (section “Introduction”), most RCMs developed and in use today still include subsets of physics parameterizations that are adapted from their host GCMs. Driven by high performance computing and the need to improve accuracy, both global and regional climate models are including more and more sophisticated parameterizations for all physical processes, which together with increasing model resolution, demand significant advances in high performance computing to support climate modeling research.

Modeling Issues

The climate of a region is determined by the large-scale atmospheric circulation as well as regional forcings such as topography within the region, and how they interact through various physical and dynamical processes. For example, the regional climate of the US Great Plains is strongly influenced by atmospheric circulation that brings moisture from the Gulf of Mexico during summer. How much precipitation is produced over land depends on moisture convergence, which is influenced not only by large-scale circulation patterns, but mesoscale features such as the Great Plain Low Level Jet, propagating disturbance from the Rocky Mountain, and local moisture sources from the land surface also play an important role. Therefore in the nested regional climate modeling approach, regional climate simulations depend on both the lateral boundary conditions that control the large-scale circulation, regional topography and land cover/land use features being resolved by the model, as well as physics parameterizations that ultimately determine the local changes in the energy, moisture, and momentum as influenced by the large-scale circulation and regional forcings.

Because of the dependence on large-scale circulation, large biases in global climate simulations used to provide lateral boundary conditions could have detrimental effects on the regional climate simulations under the nesting approach. Even if the global climate simulations were perfect, the lateral boundary conditions do not uniquely define the regional climate because the associated boundary value problem (i.e., solving the hyperbolic equations) is ill posed. Relaxation methods such as proposed by Davies [8] convert the hyperbolic equations to the well-posed parabolic form. However, mismatches between the large-scale circulation simulated by the regional models and the imposed atmospheric states at the lateral boundaries that may result from differences in grid resolution, physics, and dynamical formulations between the global and regional models can induce errors that propagate to the interior of the domains and contaminate the regional simulations [56]. This issue also leads to the sensitivity of the simulated regional climate to the domain size and locations of the lateral boundaries – an undesirable feature as it introduces uncertainties to the simulation results.

To address the validity of the nested regional climate modeling approach, a series of idealized numerical experiments have been designed and performed to assess the various assumptions used in regional climate modeling. The idealized experimental framework, known as "Big Brother Experiments (BBE)" [10], addresses modeling issues specifically related to the nested regional climate modeling approach. The Big Brother Experiment protocol consists of performing a high-resolution global climate simulation, referred as the Big Brother, BB, that serves as reference against which a regional climate simulation, referred as the Little Brother, LB, would be compared (Fig. 3). The BB, with proper spatial filtering to remove the fine scales to emulate coarse resolution global climate simulations, provides lateral boundary conditions for driving the LB. The differences between the climate simulated by the LB and the reference BB could be attributed to the nesting approach of the



Regional Climate Models. Figure 3
Flow chart of the Big Brother Experiment (BBE). The high-resolution large domain RCM simulation is used as a virtual reality to evaluate the high-resolution simulation generated by the same RCM for a smaller domain achieved through nesting (Source Laprise et al. [31] © 2008 *Meteor. Atmos. Phys.*)

limited-area model. Unlike the evaluation of real-world simulations that depends on the fidelity of model physics and availability of observational data, the idealized BBE framework allows different nesting-specific issues (e.g., the relaxation treatment and width of the buffer zone, frequency of LBC update) to be evaluated regardless of limitations of model physics and data because deficiencies of the nesting approach can be identified and quantified based on the comparison of the LB and BB alone.

A series of studies using the BBE protocol has been performed, focusing on different modeling issues specific to the nested regional modeling approach. As summarized by Laprise et al. [31], the BBE shows that the LB is capable of generating small-scale features absent from the lateral boundary conditions, and the small-scale features are consistent with the BB. These results demonstrate that the nested regional climate modeling approach does work as designed. That is, given large-scale conditions provided by the GCMs at the lateral boundaries, the RCMs can downscale to produce finer scale features absent from the GCMs. Moreover, the fine scales produced by the RCMs are consistent with what the GCMs would generate if they were applied at similar spatial resolution as the regional models. However, the small scales are not uniquely defined by the lateral boundary conditions and the domain-specific regional forcings, as the interactions between the two can be sensitive to small perturbations in the initial conditions that alter the time evolution of the small scales. The variations produced in regional simulations by perturbations in the initial conditions have been called "internal variability," as they relate to internal processes rather than external or LBC forcings. This issue has also been investigated by others (e.g., [3, 9, 21, 27]) who found that model internal variability depends on factors such as seasons, atmospheric flow regimes, and domain size. This puts a caveat on using single member short (seasonal and sub-seasonal) RCM simulations for model evaluation or hypothesis testing, as internal variability may overwhelm the signals (e.g., model errors or model response to external perturbations) being sought.

To address the issue of internal variability, ensemble modeling with perturbed initial conditions can be performed to quantify the internal variability and its impacts on model errors or model response. Alternatively, different techniques have been developed to constrain the large scales simulated by the regional models by the global climate simulations or global analyses throughout the regional domains. Spectral nudging [2, 28, 44, 54] is one example of such techniques. With spectral nudging, both the regional climate simulation and the global analyses or global climate simulations that provide lateral boundary conditions are decomposed into different spectral components in space. The simulated large-scale spectral components are nudged toward that of the global data using relaxation to provide stronger large-scale constraints on the regional climate simulations than that imposed by the lateral boundary conditions alone. These methods reduce the mismatch between the simulated large scales and the imposed lateral boundary conditions that contaminate the regional simulations. They also reduce internal variability, so simulation with a single initial atmospheric condition may be sufficient to assess model errors or estimate model response to external forcings. On the other hand, the degree of nudging to be applied to constrain the large scales can be rather arbitrary. Also, one may argue that by nudging the large scales of the regional climate simulations toward the global climate simulations, these methods increase the dependence of the regional simulations on the skill of the global models and eliminate the potential for the regional models to improve the large scales through upscaling of mesoscale features that are better resolved by the regional models [43].

Besides some form of large-scale nudging applied throughout the regional model domain, some studies have proposed a different mode of simulating regional climate by applying regional climate models with frequent initialization of the atmosphere to simulate short time segments that are then concatenated to compose the long-term regional climate simulations [45, 47]. This method takes advantage of the time period of limited error growth shortly after model initialization so the mismatch between the simulated and imposed large scale is small even without additional constraints on the large scale in the interior of the model domain. Two-way nesting of global and regional models has also been proposed as an approach to reduce large-scale inconsistency that may develop in one-way nested regional climate models because the upscaled influences of the regional models are included in the global models through feedbacks [40]. Such an approach has only been evaluated in a few studies [26, 41], and the results have been encouraging.

In summary, although the limited-area or nested modeling approach upon which regional climate models are based is an ill-posed boundary value problem, practical solutions such as the relaxation boundary treatment and spectral nudging of the large scale throughout the regional climate model domain have been developed and found to work well for a large number of cases. Furthermore, idealized experiments have confirmed most of the assumptions used in regional climate modeling [31]. However, uncertainty in regional climate simulations remain, owing in part to issues such as physics parameterizations, model resolutions, and initial conditions that are common to both global and regional climate models, and issues such as dependence on the lateral boundary condition, boundary treatment, regional domain size and location, and use of interior nudging that are specific to the nested regional modeling. Reviews and discussions of these issues can be found in Giorgi and Mearns [16], Laprise et al. [31], Leung et al. [35], and Wang et al. [55]. More research is needed to better understand the sensitivity of regional climate simulations to different factors and develop ways to reduce the uncertainty introduced by the nested modeling framework.

Evaluating Regional Climate Models

Model evaluation is important for assessing and documenting model skill and how it may evolve over time as changes and improvements are added to the models. It also provides information needed to understand model behaviors and diagnose model biases, and to assess uncertainties associated with the regional climate simulations. Model evaluation is achieved primarily by comparing model simulations with observations. The most common observation data used in evaluating regional climate simulations are atmospheric data such as 500 hPa height and upper level winds from global analyses, and surface temperature and precipitation from surface meteorological stations (e.g., Climatic Research Unit (CRU) and University of Delaware (UD) datasets), satellite-derived data (e.g., Tropical Rainfall Measurement Mission (TRMM)), and integrated station/satellite products (e.g., Global Precipitation Climatology Project (GPCP) [24] and Climate Prediction Center (CPC) Merged Analysis of Precipitation CMAP) [59]. These data are typically spatially interpolated to uniform latitude/longitude grids.

Both surface temperature and precipitation have high spatial variability due to surface topography and other factors. The effect of topography is relatively easy to account for in surface temperature as it varies with altitude more or less according to the standard temperature lapse rate, but its influence on precipitation is more spatially variable depending on a number of factors such as wind direction and surface slope and aspect. Statistical methods such as Parameter-elevation Regressions on Independent Slopes Model (PRISM) [7] have been developed to account for surface topographical effects in gridded precipitation data. There is a continuing need to develop high temporal and spatial resolution datasets for evaluating regional climate models. Recent efforts in Europe [22] and Asia [60] have made great strides in providing high resolution (0.1° and 25/50 km resolution for Europe and 0.25° and 0.5° for Asia) gridded daily precipitation data for model evaluation and analysis, although differences among datasets can still be substantial in mountain areas due to measurement methods, retrieval algorithms, grid resolutions, and whether topographic effects are explicitly accounted for.

By comparing observed and simulated surface temperature, precipitation, and atmospheric fields, model biases can be identified. However, determining the sources of model errors and thereby providing guidance on reducing model biases requires more information. Observations that can be used to diagnose model errors are more limited. For example, to understand model biases in surface temperature, it is useful to know which components (e.g., net shortwave and longwave radiation and sensible and latent heat fluxes) of the surface energy budgets may be in error. Ground-based measurements of the surface energy fluxes are limited both spatially and temporally. However, some flux data are available from a global network (FLUXNET) of about 400 micrometeorological tower sites that provide continuous measurements, some dating back to 1996. There is a challenge in relating point measurements of surface fluxes with model simulations that represent grid box averages. Satellite retrievals of radiation fluxes are available globally for recent decades, but large differences exist among different datasets such as Clouds and the Earth's Radiant Energy System (CERES) and International Satellite Cloud Climatology Project (ISCCP).

Diagnosing errors in precipitation is even more challenging because precipitation is the end product of many interactive processes. Although precipitation is more directly related to clouds, measurements of cloud macrophysical and microphysical properties are limited. Cloud climatologies are available from ISCCP and CERES, but the grid resolution is relatively coarse (280 km for ISCCP and 1° for CERES) compared to regional models. Furthermore, errors in simulating clouds may be reflecting other problems because myriads of processes can influence the formation and evolution of clouds. Higher temporal and spatial resolution precipitation can provide a means to evaluate temporal variability from diurnal to seasonal, and probability distribution of precipitation rates, which can provide important clues to processes that may not be well represented in models. Some surface hydrological variables such as river runoff and snowpack may also be used to infer model biases in precipitation or combinations of precipitation and temperature biases.

Besides advances in the development of datasets for model evaluation, the methods used to evaluate models have also become more sophisticated. In the 1990s, comparisons of observations and model simulations were mostly limited to seasonal/annual and regional averages, but more studies now also compare observed and simulated temporal and spatial variability such as interannual variability and spatial distributions. With more studies producing longer regional climate simulations, more aspects of the simulations such as diurnal variability, extreme statistics, regime-specific features, frequency distributions, co-variability of different variables (e.g., between temperature and precipitation), and parameters that reflect the strengths of feedback processes have been evaluated (e.g., comparing land-atmosphere coupling strengths between models).

Although model evaluation studies are broadly aimed at understanding and quantifying model biases so model improvements can be made, some efforts also evaluate specific aspects such as precipitation and runoff [32], wind resources [57] of the regional simulations to provide practical guidance on their usefulness to provide climate information for impact assessments and resource management or planning. To support more detailed analyses, the requirements on model outputs have significantly increased as higher temporal frequency model outputs (e.g., hourly and daily) and more simulated state variables and tendencies are becoming more commonly archived.

Besides comparing model simulations with observations, model intercomparison can add significant information to understand and characterize model differences and uncertainties. The Atmospheric Model Intercomparison (AMIP) project [15] was initiated in the early 1990s to determine the systematic errors of global atmospheric models used to simulate long-term climate. Since the first AMIP project, many intercomparison projects have been developed to evaluate climate models used in different simulation modes. Similar coordinated projects have also been initiated to intercompare regional climate simulations since the mid-1990s. The first of such projects is the Project to Intercompare Regional Climate Simulations (PIRCS) [53]. PIRCS includes two phases, with the first phase focusing on simulations of two anomalous years, the 1988 drought and 1993 flood in the US Great Plains, and the second phase comparing multiyear simulations over North America. All simulations were driven by global reanalysis data and observed sea surface temperature. Besides regional climate models, one global stretched-grid model also participated in PIRCS for comparison between two dynamical frameworks for regional climate modeling. Following PIRCS, several intercomparison projects were developed to compare regional climate simulations over the Arctic (ARCMIP) (http://curry.eas.gatech.edu/ARCMIP/) and East Asia (RMIP) [14]. More discussions of intercomparison projects that focus on climate change simulations are provided in section "Dynamical Downscaling."

Applications of Regional Climate Models

Climate Process Studies

An important application of regional climate models is to advance the understanding of regional climate processes. In this context, regional climate models are often used to test hypotheses of how different regional forcings or feedback mechanisms play a role in regional climate variability and change.

For example, Leung et al. [34] used long-term simulations of the western USA to investigate the role of topography on precipitation spatial distribution during El-Nino and La-Nina events. Comparing precipitation during El-Nino years with the 20-year simulated climatology, they found a positive-negative-positive anomaly pattern in the Olympic Mountains and the west side and east side of the Cascades Mountains in the US Pacific Northwest. The pattern was found to be a result of the interactions between the large-scale atmospheric circulation that are influenced by the ENSO conditions and the orientation of the mountains. With atmospheric flow assuming a more southwesterly rather than a westerly direction during El-Nino years, the rain shadow created by the north-south oriented Cascades Mountains is reduced, resulting in more precipitation reaching the lee side of the mountains. Such regional anomaly patterns are generally not found in global reanalyses or global climate simulations because of their coarser resolution, but are consistent with observed precipitation and streamflow anomalies in the region.

Hughes and Hall [25] performed regional climate simulations for the western USA to investigate large-scale and local controls on Santa Ana winds in Southern California. Using a simulation at 6 km resolution, their analysis showed that both large-scale anomaly corresponding to a high pressure over the Great Basins, and local thermodynamic forcing due to surface temperature gradient between the cold desert (Mojave Desert) and warm ocean create pressure gradients that drive offshore winds. The latter was found to be particularly important in determining the timing of Santa Ana winds, which occur more frequently during December when the temperature gradient between the desert and Pacific coast is the largest.

The role of soot on mountain snowpack and hydrology was investigated by Qian et al. [48] using regional climate simulations with and without soot deposition in western USA. Their study shows that soot-induced snow-albedo perturbations increase the surface net solar radiation flux during late winter to early spring. This increases the surface air temperature and reduces snow accumulation and spring snowmelt, causing a trend toward earlier snowmelt. Snow-albedo feedback was found to play an important role in amplifying the soot effects in the mountains.

Riddle and Cook [50] used regional climate simulations to study the mechanism of abrupt rainfall transition in the Greater Horn of Africa. The yearly monsoon jump of about 20° latitude during April and May was found to coincide with abrupt circulation changes associated with the Somali jet that develops during that time. The cross-equatorial branch of the Somali jet brings moisture to the southern slopes of the Ethiopian plateau, which then produces the abrupt rainfall transition in the region.

To investigate why temperature in the central USA has cooled by 0.2–0.8°C in the late twentieth century, instead of warmed as in most continental regions, Pan et al. [46] used a regional climate model and found that under a global warming scenario, increased moisture from the Gulf of Mexico due to warming and increasing occurrence of the Great Plain Low Level Jet (LLJ) in the south and decreasing occurrence in the north enhances atmospheric moisture convergence and cloudiness and precipitation in the central USA. These changes replenish soil moisture during summer, which increases late-summer evapotranspiration and suppresses daytime maximum temperature, and hence the "warming hole." Because of coarse resolution, most GCMs cannot simulate the observed "warming hole" in the late twentieth century.

Regional climate models also offer great potentials to understand the mechanisms of extreme events and their projected changes in the future. For example, Seneviratne et al. (2006) performed two regional climate simulations with and without land-atmosphere interactions to investigate the role of land-atmosphere feedbacks on heat waves in Europe. They showed that soil moisture – temperature and soil moisture – precipitation feedbacks increase summer temperature variability in central and eastern Europe. Under climate change, the region of stronger land-atmosphere coupling shifts northward in response to greenhouse warming to central and eastern Europe, and enhances summer temperature variability and increases the potential for more frequent heat waves in that region.

In the above examples, high resolution is important for the model to reproduce the observed climatology of temperature, precipitation, wind, or snowpack, which in the western USA, Europe, and the Greater Horn of

Africa depend strongly on regional orography. High resolution is also important for simulating soot deposition caused by anthropogenic emissions in cities being carried to the mountains downwind, or LLJ and its effects on cloudiness and precipitation. Successful simulations of the base states and model ability to simulate regional forcings and feedback mechanisms (e.g., snow-albedo, soil moisture – temperature feedbacks, LLJ – precipitation coupling) are critical for assessing their role in the observed regional climate phenomena.

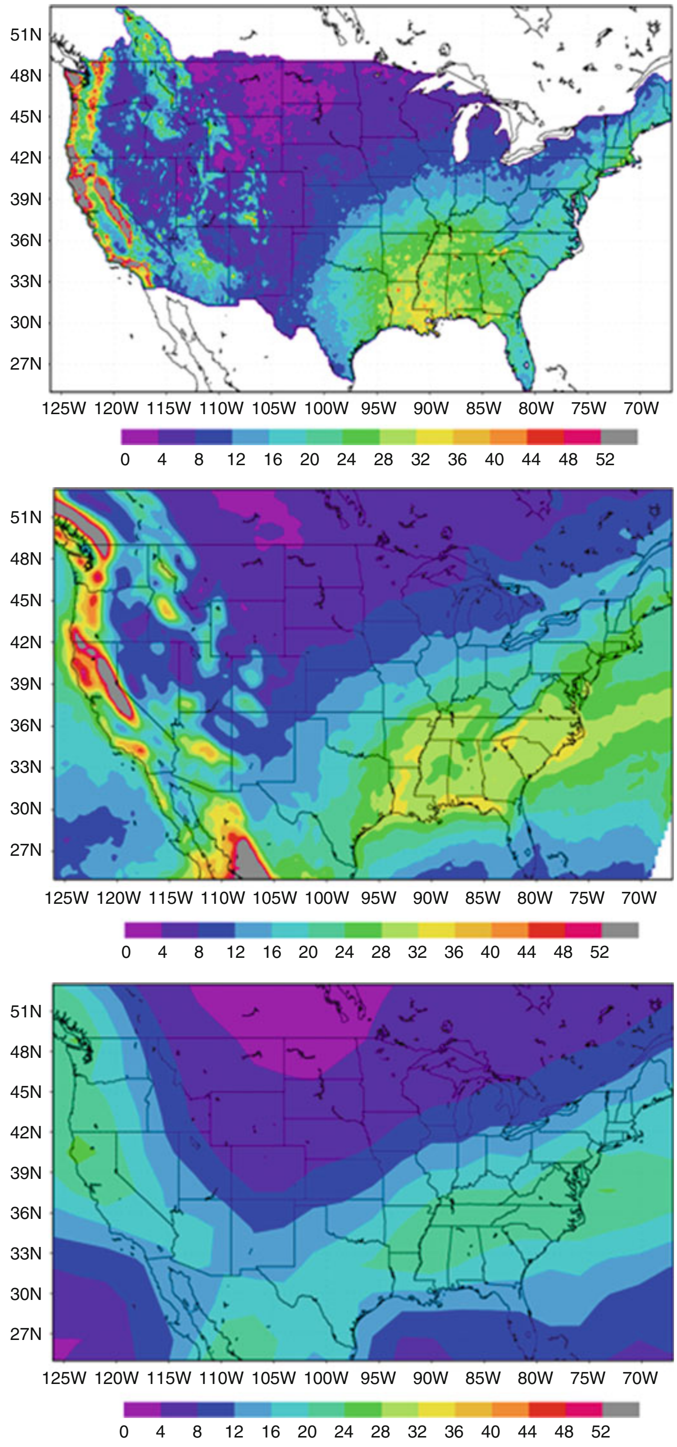
Dynamical Downscaling

Dynamical downscaling is an important application of regional climate models, which aims to provide more spatially resolved climate predictions or projections provided by GCMs. Most of the downscaling applications to date are related to climate change projections. Early efforts described the use of an individual RCM to dynamically downscale climate change projections by a specific GCM. Typically only a single emission scenario such as the business-as-usual scenario (1% increase of CO_2 per decade) was used. Although GCMs generally produce simulations that cover preindustrial to 2100, RCM simulations are usually performed only for two time segments of 10–30 years corresponding to a current and a future time period.

Giorgi et al. [18] reported the first set of studies on using a regional climate model to dynamical downscale climate change scenario for Europe and the western Mediterranean basin. The GCM and RCM they used had a spatial resolution of R15 (about 400 km) and 70 km, respectively. The current and future climate corresponds to the equilibrium conditions simulated by the GCM using 1 × CO_2 (preindustrial level) and 2 × CO_2 (doubling of preindustrial level), respectively. Although the GCM generally reproduced the basic seasonal migration patterns of storm tracks, significant biases were also found in large-scale features such as the location and strength of the North Atlantic jet, cold tropospheric temperature and low tropospheric relative humidity, and underprediction of summer precipitation. Overall, the RCM was found to inherit most of the large-scale biases from the GCM, but the spatial distribution of temperature and precipitation was better simulated due to topographic effects. In addition, the RCM produced more spatially refined temperature and precipitation change scenarios. The RCM also simulated significant sub-GCM-scale changes in surface hydrological variables such as snow depth and runoff.

Following a similar approach, Leung and Ghan [36] used a regional climate model driven by GCM 1 × CO_2 and 2 × CO_2 simulations to produce climate change scenarios for the western USA. However, much more spatially resolved simulations of temperature and precipitation were produced by using a subgrid parameterization of orographic precipitation and vegetation [38,39]. This method divides a model grid cell into subgrid surface elevation and vegetation classes based on high resolution (1 km) DEM and vegetation data. The influence of topography and vegetation on atmospheric and land surface processes is represented through a parameterization that accounts for orographic effects on clouds, which then affect precipitation and surface hydrology. During postprocessing, surface temperature and precipitation, among other variables, simulated for each subgrid class are mapped geographically to 1 km resolution based on the DEM and vegetation data. This approach greatly improves the simulation of surface temperature, precipitation, and snowpack compared to the GCMs. Their results show that snowpack will potentially be reduced by up to 50% under a 2 × CO_2 scenario. They also found a strong elevation dependence of climate change signals in temperature, precipitation (amount and phase), snow cover, and runoff (see also [1,20] for a discussion of elevation dependence of climate change signals in mountainous regions).

In the 2000s, as more GCM transient simulations became available and the regional modeling community has grown, more studies have been published that investigated the potential effects of climate change in different climatic regimes or geographical locations. Figure 4 shows an example of cold season heavy precipitation (95th percentile) simulated by a GCM and an RCM driven by the GCM boundary conditions, using the same models described by Leung et al. [32], except for a change in the regional domain to cover the conterminous USA. Comparison of the simulated and observed heavy precipitation shows that the RCM reproduced the observed spatial distribution of heavy precipitation better than the GCM primarily because of the increased spatial resolution. As regional climate



Regional Climate Models. Figure 4
An example of cold season heavy (95th percentile) precipitation simulated by a GCM (*Top*) and an RCM (*middle*) and comparison with observation (*Bottom*) over the USA. The prominent effects of topography are well captured by the RCM at 36 km grid resolution compared to the GCM, which was applied at roughly 250 km resolution

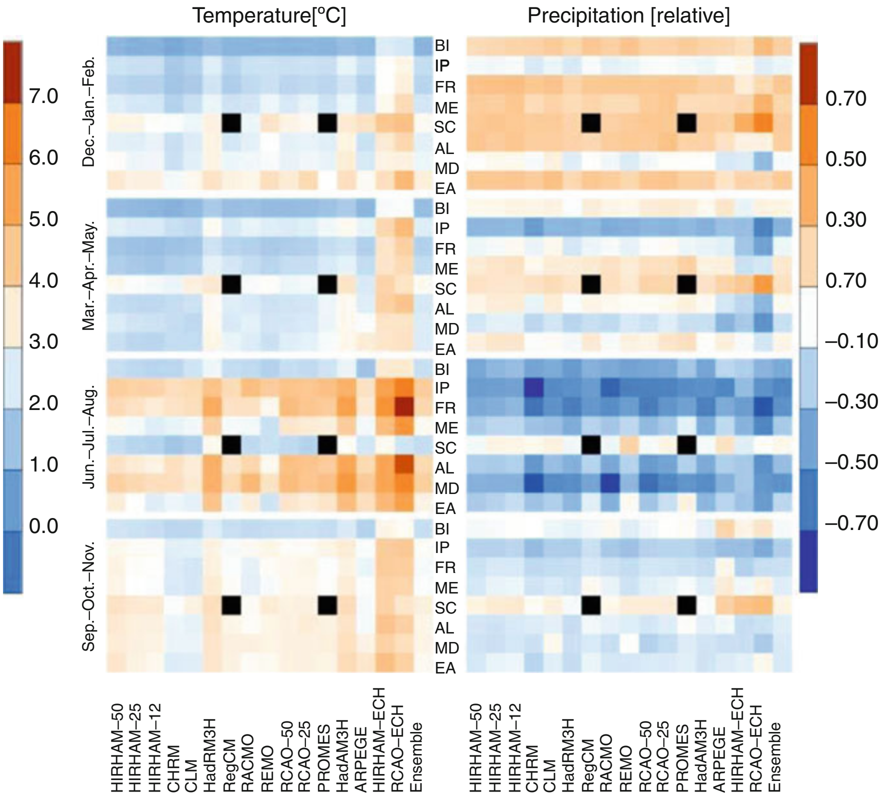
information is useful for assessing climate impacts and addressing climate adaptation, many studies that involve the use of regional climate models for producing regional climate change scenarios included scientists and stakeholders of the specific regions being studied to focus on subjects both scientifically interesting and societally relevant. The regional human resources and knowledge base that have been tapped have proven beneficial and contributed to more diverse analyses and applications of the climate change results. More examples of these efforts have been summarized in Christensen et al. [5].

Besides individual efforts of using a particular RCM to downscale climate predictions or projections from a particular GCM, larger efforts have also been coordinated to develop ensembles of dynamically downscaled simulations. Common to these coordinated efforts is the objective to fill the gap in providing regional climate change scenarios for different geographic regions and to improve the characterization of uncertainty of the scenarios. To this end, an ensemble modeling approach is used in which multiple RCMs are nested within multiple GCMs to generate a matrix of regional climate change scenarios to facilitate the interpretation and characterization of uncertainty of regional climate change. These efforts also enable large, multi-model datasets to be archived following common protocols similar to the AMIP and CMIP efforts adopted by the GCM community over the last two decades.

In Europe, two large coordinated projects, PRUDENCE [5] and ENSEMBLES [23], intercompared regional climate models driven by global reanalysis as well as global climate simulations for the current and future climate. PRUDENCE designed, executed, analyzed, and synthesized regional climate scenario development for Europe. In brief, four GCMs and ten RCMs were involved to produce regional climate scenarios at 50 km resolution, but a few scenarios at 20 km resolution were also produced. Two time slices were simulated by each RCM, corresponding to 30 years of control and future (2071–2100) conditions. Two emission scenarios, A2 and B2, were considered, and some GCMs and RCMs provided multiple ensemble members (using different initial conditions) for assessing internal variability. Although only 28 combinations out of the full matrix of GCM, RCM, and scenario combinations were performed, PRUDENCE provided sufficient model outputs to evaluate the variance due to the four sources of uncertainty: GCM, RCM, scenario, and sampling. Figure 5 summarizes the surface temperature and precipitation changes simulated by the regional models for Europe.

One of the main conclusions of PRUDENCE is that the largest source of uncertainty resides in the GCM boundary conditions applied to the RCM [11]. The choice of RCM becomes more important, however, for certain subregions or seasons (summer in particular). Furthermore, many local features and aspects of extremes can vary substantially between RCMs [30] to alter the climate change signal from that simulated by the driving GCM. For example, RCM simulations performed at higher resolution (12 km vs 25 km) reduced the magnitude of future summer drying over southern Europe [4]. This effect could be attributed to the diminished control of the LBCs on RCM simulations during summer, and the general tendency of RCM to produce more precipitation at higher resolutions (e.g., [33,49]).

Building on the foundation of PRUDENCE, ENSEMBLES is the largest and most comprehensive RCM comparison project conducted to date. Focusing again on Europe, ENSEMBLES utilized 15 GCMs and 11 RCMs to create a large GCM-RCM matrix for a single emission scenario (A1B). Simulations were also performed with reanalysis boundary conditions at 25 km and 50 km horizontal resolution. Interestingly, higher resolution (25 km vs 50 km) did not improve the simulation of large-scale weather types [51] or seasonal precipitation [49] by many RCMs, suggesting that physics and/or downscaling approach (i.e., dynamical framework) may be more important than resolution. The most novel aspect of the ENSEMBLES project is the construction and use of a set of metrics to weight models according to their performance to construct an ensemble mean [6]. However, application of the weights to the GCM-forced RCM simulations for the twentieth century did not substantially improve the performance of the multi-model temperature or precipitation mean over the unweighted multi-model mean when averaged over Europe. This suggests that more research is needed to further explore the productive use of ensembles of climate change scenarios to reduce uncertainty.



Regional Climate Models. Figure 5

An overview of seasonal changes in surface temperature (degree C) (*left*) and precipitation (relative change) (*right*) simulated by the PRUDENCE regional climate models for different analysis areas (row) and models (column). The analysis areas are: *BI* British Isles, *IP* Iberian Peninsula, *FR* France, *ME* Mid-Europe, *SC* Scandinavia, *AL* Alps, *MD* Mediterranean, *EA* Eastern Europe. Results from 17 regional simulations (some are produced by the same model at different resolutions) are shown, but some simulations did not cover certain geographical areas (shown by the *black squares*) (Source Christensen and Christensen [4] © 2007 *Climatic Change*)

The North American Regional Climate Change Assessment Program (NARCCAP) [42] is another coordinated project similar to PRUDENCE and ENSEMBLES, but with a geographic focus on North America. The NARCCAP GCM-RCM matrix includes mapping 4 GCMs with 6 RCMs statistically for a more balanced design for uncertainty analysis. In addition, two high-resolution time-slice global simulations are included for comparison with the RCM simulations over North America. More recently, CORDEX (a COordinated Regional Downscaling EXperiment) has been developed to coordinate regional climate change scenario development for all continents around the world, and to foster international collaborations and promote interactions and communications between the various communities involved in scenario development and applications. The CORDEX design is similar to the multi-GCM/RCM matrix used in PRUDENCE, ENSEMBLES, and NARCCAP, but an additional level of uncertainty being assessed is model dependence on climate regimes and/or geographic locations. Therefore, an important CORDEX effort is to develop and compare climate simulations across different continents. Additionally, CORDEX encourages the development of Regional Analysis and Evaluation Teams to develop a set of regionally specific metrics for model evaluation, collect observational data, design experiments to investigate the added value of RCMs, and evaluate the ensemble of simulations from CORDEX.

Besides climate change, dynamical downscaling has also been applied to the area of seasonal climate predictions, but to a much lesser extent compared to downscaling of climate change simulations. The Multi-Regional Ensemble Downscaling (MRED) is a coordinated project in which multiple RCMs were used to downscale global seasonal climate forecasts for the USA (http://ecpc.ucsd.edu/projects/MRED/). Dynamical downscaling has also been used to develop regional analysis for studying climate variability and trends. Unlike regional analysis such as the North American Regional Reanalysis that assimilates observation data in regional models driven by global analysis, the dynamical downscaling approach assimilates only global analysis, but no additional observational data within the regional model domains to generate regional climate information. As examples, Sotillo et al. [52] used a regional model to downscale global reanalysis to generate a high-resolution 44-year atmospheric analysis for the Mediterranean Basin. Kanamitsu and Kanamaru [29] used a regional climate model at 10 km resolution driven by a global reanalysis in the California Reanalysis Downscaling at 10 km (CaRD10) project to produce 57 years of regional analysis for California.

Although numerous studies that evaluated different aspects of regional climate simulations using observations have demonstrated some skill in simulating regional temperature and precipitation, the skill depends very much on the large-scale data used to drive the model, the model physics, and how the models were configured. More recently, besides asking what ability RCMs have in reproducing observed climate features, the question of whether dynamical downscaling adds values to global climate simulations has become an important topic. Essentially, this begs the question of whether the additional step of running regional climate models as a means to dynamically downscale global climate simulations indeed provides additional (useful) information not available from the global climate simulations. One way to address this question is to define and apply various metrics to quantitatively measure the added skill or added information provided by the regional models. For example, spatial filters can be used to partition the model-simulated variability to a larger scale that is resolved by the global simulation and a smaller scale that is beyond the limit resolved by the global simulation. The amount of finer scale variability generated by the regional models is considered value added since it provides climate information beyond what the global simulations provide (e.g., [12]). Similarly, spectral decomposition can be applied to simulated quantities such as different components of the surface water budgets and forecast skill to determine the added value of regional modeling.

Another aspect of evaluating the value added by RCMs is to compare dynamical downscaling with statistical downscaling, which is computationally a much cheaper method to produce regional climate information. To date, comparison of dynamical and statistical downscaling methods is limited to a few studies. Wood et al. [58] represent an early effort to apply a simple statistical downscaling method called Bias Correction Spatial Disaggregation (BCSD) to not only global simulations, but also regional climate simulations. The latter is a hybrid approach that combines dynamical and statistical downscaling. Comparing statistically downscaled simulations driven by global and regional simulations with the global and regional simulations, this study showed that hydrologic response to climate change can be enhanced using the hybrid approach compared to applying statistical downscaling directly to the GCM outputs because the RCM simulated larger warming in mountainous areas as a result of snow-albedo feedbacks, which are not captured by GCM or statistical downscaling.

Future Directions

In summary, both idealized experiments and real case applications have demonstrated that nested regional climate modeling is a viable approach for regional climate simulations. However, applications of regional climate models must be exercised with care because many factors can introduce uncertainty in the simulated results. These factors, which include domain size and location, physics parameterization, model resolution, lateral boundary condition and treatment, and use of interior nudging, must be carefully assessed before proceeding to long-term climate simulations. More research is also needed to better understand the sensitivity of regional climate simulations to those factors and develop ways to characterize and reduce uncertainty introduced by the nested modeling framework. As computing resources allow global models to be applied at higher and higher spatial resolution, and alternative approaches such as global variable resolution models become feasible, more research is needed to evaluate and compare different approaches to modeling regional climate to establish their validity and usefulness in addressing different aspects of climate research and applications.

Acknowledgments

I would like to thank my colleagues at the Pacific Northwest National Laboratory and my collaborators over the years, whose research has inspired me. I also thank them for sharing their ideas, knowledge, and results with me.

Bibliography

1. Beniston M, Diaz HF, Bradley RS (1997) Climatic change at high elevation sites: an overview. Clim Chang 36:233–251
2. Castro CL, Pielke RA Sr, Leoncini G (2005) Dynamical downscaling: an assessment of value added using a regional climate model. J Geophys Res 110. doi:10.1029/2004JD004721, D05108
3. Caya D, Biner S (2004) Internal variability of RCM simulations over an annual cycle. Clim Dyn 22(1):33–46
4. Christensen JH, Christensen OB (2007) A summary of the PRUDENCE model projections of changes in European climate by the end of this century. Clim Chang 81:7–30. doi:10.1007/s10584-006-9210-7
5. Christensen JH, et al (2007) Regional climate projections. In: Solomon S, Qin D, Manning M, Chen Z, Marquis M, Avery KB, Tignor M, Miller HL (eds) Climate change 2007: the physical science basis, contribution of Working Group I to the Fourth Assessment Report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge, UK/New York
6. Christensen JH, Rummukainen M, Lenderink G (2009) Formulation of very high-resolution regional climate model ensembles for Europe, chapter 5. In: van der Linden P, Mitchell JFB (eds) ENSEMBLES: climate change and its impacts: summary of research and results from the ENSEMBLES project. Met Office Hadley Centre, Exeter, 160pp
7. Daly C, Neilson RP, Phillips DL (1994) A statistical-topographic model for mapping climatological precipitation over mountanious terrain. J Appl Meteor 33:140–158
8. Davies HC (1976) A lateral boundary formulation for multi-level prediction models. Quart J Roy Meteor Soc 102:405–418
9. de Elía R, Laprise R, Denis B (2002) Forecasting skill limits of nested, limited-area models: a perfect-model approach. Mon Weather Rev 130:2006–2023
10. Denis B, Laprise R, Caya D, Côté J (2002) Downscaling ability of one-way-nested regional climate models: the Big-brother experiment. Clim Dyn 18:627–646
11. Déqué M, Rowell DP, Lüthi D, Giorgi F, Christensen JH, Rockel B, Jacob D, Kjellström E, Castro M, van den Hurk B (2007) An intercomparison of regional climate simulations for Europe: assessing uncertainties in model projections. Clim Chang 81:53–70
12. Di Luca A, de Elía R, Laprise R (2011) Assessment of the potential added value in multi-RCM simulated precipitation. Clim Dyn. doi:10.1007/s00382-011-1068-3
13. Dickinson RE, Errico RM, Giorgi F, Bates GT (1989) A regional climate model for the western United States. Clim Chang 15:383–422
14. Fu C, Wang S, Xiong Z, Gutowski WJ, Lee D-K, McGregor JL, Sato Y, Kato Hi, Kim J-W, Suh M-S (2005) Regional climate model intercomparison project for Asia. Bull Amer Meteorol Soc 86. doi:10.1175/BAMS-86-2-257
15. Gates WL (1992) AMIP: the atmospheric model intercomparison project. Bull Amer Meteorol Soc 73:1962–1970
16. Giorgi F, Mearns LO (1999) Introduction to special section – regional climate modeling revisited. J Geophys Res 104(D6):6335–6352
17. Giorgi F, Bates GT (1989) On the climatological skill of a regional model over complex terrain. Mon Weather Rev 117:2325–2347
18. Giorgi F, Marinucci MR, Visconti G (1990) Use of a limited area model nested in a general circulation model for region climate simulation over Europe. J Geohys Res 95:18,413–18,431
19. Giorgi F, Marinucci MR, DeCanio G, Bates GT (1993) Development of a second generation regional climate model (REGCM2): cumulus cloud and assimilation of lateral boundary conditions. Mon Weather Rev 121:2814–2832
20. Giorgi F, Hurrell JW, Marinucci MR, Beniston M (1997) Elevation signal in surface climate change: a model study. J Clim 10:288–296
21. Giorgi F, Bi X (2000) A study of internal variability of a regional climate model. J Geophys Res 105:29503–29521
22. Haylock MR, Hofstra N, Klein Tank AMG, Klok EJ, Jones PD, New M (2008) A European daily high-resolution gridded dataset of surface temperature and precipitation. J Geophys Res (Atmos) 113(D20119). doi:10.1029/2008JD10201
23. Hewitt CD, Griggs DJ (2004) Ensembles-based predictions of climate changes and their impacts. EOS 85:566
24. Huffman GJ, Adler RF, Morrissey M, Bolvin DT, Curtis S, Joyce R, McGavock B, Susskind J (2001) Global precipitation at one-degree daily resolution from multisatellite observations. J Hydrometeor 2:36–50
25. Hughes M, Hall A (2010) Local and synoptic mechanisms causing Southern California's Santa Ana winds. Clim Dyn 34. doi:10.1007/s00382-009-0650-4
26. Inatsu M, Kimoto M (2009) A scale interaction study on East Asian cyclogenesis using a general circulation model coupled with an interactively nested regional model. Mon Weather Rev 137. doi:10.1175/2009MWR2825.1

27. Ji YM, Vernekar AD (1997) Simulation of the Asian summer monsoons of 1987 and 1988 with a regional model nested in a global GCM. J Climate 10:1965–1979
28. Kanamaru H, Kanamitsu M (2007) Scale-selective bias correction in a downscaling of global analysis using a regional model. Mon Weather Rev 135:334–350
29. Kanamitsu M, Kanamaru H (2007) 57-Year California reanalysis downscaling at 10 km (CaRD10) part 1. System detail and validation with observations. J Climate 20:5527–5552
30. Kjellström E, Bärring L, Jacob D, Jones R, Lenderink G, Schär C (2007) Modelling daily temperature extremes: recent climate and future changes over Europe. Clim Chang 81:249–265
31. Laprise R, de Elía R, Caya D, Biner S, Lucas-Picher Ph, Diaconescu EP, Leduc M, Alexandru A and Separovic L (2008) Challenging some tenets of regional climate modelling. Meteor. Atmos Phys 100, Special Issue on Regional Climate Studies, 3–22. doi:10.1007/s00703-008-0292-9
32. Leung LR, Qian Y, Bian X, Washington WM, Han J, Roads JO (2004) Mid-century ensemble regional climate change scenarios for the western United States. Clim Chang 62(1–3):75–113
33. Leung LR, Qian Y (2003) The sensitivity of precipitation and snowpack simulations to model resolution via nesting in regions of complex terrain. J Hydrometeorol 4(6):1025–1043
34. Leung LR, Qian Y, Bian X, Hunt A (2003) Hydroclimate of the western United States based on observations and regional climate simulation of 1981–2000. Part II: mesoscale ENSO anomalies. J Clim 16(12):1912–1928
35. Leung LR, Mearns LO, Giorgi F, Wilby R (2003) Workshop on regional climate research: needs and opportunities. Bull Amer Meteorol Soc 84(1):89–95
36. Leung LR, Ghan SJ (1999) Pacific northwest climate sensitivity simulated by a regional climate model driven by a GCM. Part I: control simulations. J Clim 12(7):2010–2030
37. Leung LR, Ghan SJ (1999b) Pacific Northwest climate sensitivity simulated by a regional climate model driven by a GCM. Part II: 2xCO2 simulations. J Clim 12(7):2031–2053
38. Leung LR, Ghan SJ (1998) Parameterizing subgrid orographic precipitation and surface cover in climate models. Mon Weather Rev 126(12):3271–3291
39. Leung LR, Ghan SJ (1995) A subgrid parameterization of orographic precipitation. Theor Appl Climatol 52:95–118
40. Leung LR, Kuo Y-H, Tribbia J (2006) Research needs and directions of regional climate modeling using WRF and CCSM. Bull Am Meteorol Soc 87(12):1747–1751
41. Lorenz P, Jacob D (2005) Influence of regional scale information on the global circulation: a two-way nesting climate simulation. Geophys Res Lett 32:L18706. doi:10.1029/2005GL023351
42. Mearns LO, Gutowski W, Jones R, Leung R, McGinnis S, Nunes A, Qian Y (2009) A regional climate change program for North America. Eos Trans AGU 90:311–312
43. Mesinger F, Brill K, Chuang H, DiMego G, Rogers E (2002) Limited area predictability: can upscaling also take place? Research activities in atmospheric and oceanic modelling. Report No. 32, WMO/TD – No. 1105, 5.30–5.31
44. Miguez-Macho G, Stenchikov GL, Robock A (2004) Spectral nudging to eliminate the effects of domain position and geometry in regional climate model simulations. J Geophys Res 109(D13):D13104. doi:10.1029/2003JD004495
45. Pan Z, Takle E, Gutowski W, Turner R (1999) Long simulation of regional climate as a sequence of short segments. Mon Weather Rev 127:308–327
46. Pan Z, Arritt RW, Takle ES, Gutowski WJ Jr, Anderson CJ, Segal M (2004) Altered hydrologic feedback in a warming climate introduces a "warming hole". Geophys Res Lett 31: L17109. doi:10.1029/2004GL020528
47. Qian J-H, Seth A, Zebiak S (2003) Reinitialized versus continuous simulations for regional climate downscaling. Mon Weather Rev 131:2857–2874
48. Qian Y, Gustafson WI Jr, Leung LR, Ghan SJ (2009) Effects of soot-induced snow albedo change on snowpack and hydrological cycle in Western U.S. based on WRF chemistry and regional climate simulations. J Geophys Res 114:D03108. doi:10.1029/2008JD011039
49. Rauscher SA, Coppola E, Piani C, Giorgi F (2009) Resolution effects on regional climate model simulations of seasonal precipitation over Europe. Clim Dyn. doi:10.1007/s00382-009-0607-7. 28
50. Riddle EE, Cook KH (2008) Abrupt rainfall transitions over the Greater Horn of Africa: Observations and regional model simulations. J Geophys Res 113:D15109
51. Sanchez-Gomez E, Somot S, Déqué M (2008) Ability of an ensemble of regional climate models to reproduce weather regimes over Europe-Atlantic during the period 1961–2000. Clim Dyn 33(5):723–736. doi:10.1007/s00382-008-0502-7
52. Sotillo M, Ratsimandresy A, Carretero J, Bentamy A, Valero F, Gonzalez-Rouco F (2005) A high-resolution 44-year atmospheric hind-cast for the Mediterranean basin: contribution to the regional improvement of global reanalysis. Clim Dyn 25:219–236
53. Takle ES, Gutowski WJ Jr, Arritt RW, Pan Z, Anderson CJ, Silva R, Caya D, Chen S-C, Christensen JH, Hong S-Y, Juang H-MH, Katzfey JJ, Lapenta WM, Laprise R, Lopez P, McGregor J, Roads JO (1999) Project to intercompare regional climate simulations (PIRCS): description and initial results. J Geophys Res 104:19,443–19,462
54. von Storch H, Langenberg H, Feser F (2000) A spectral nudging technique for dynamical downscaling purposes. Mon Weather Rev 128:3664–3673
55. Wang Y, Leung LR, McGregor JL, Lee D-K, Wang W-C, Ding Y, Kimura F (2004) Regional climate modeling: progress, challenges, and prospects. J Meteor Soc Jpn 82(6):1599–1628
56. Warner TT, Peterson RA, Treadon RE (1997) A tutorial on lateral conditions as a basic and potentially serious limitation to regional numerical weather prediction. Bull Amer Meteor Soc 78(11):2599–2617
57. Winterfeldt J, Weisse R (2009) Assessment of value added for surface marine wind speed obtained from two regional climate models. Mon Weather Rev 137:2955–2965
58. Wood AW, Leung LR, Sridhar V, Lettenmaier DP (2004) Hydrologic implications of dynamical and statistical

approaches to downscaling climate model outputs. Clim Chang 62(1–3):189–216

59. Xie P, Yatagai A, Chen M, Hayasaka T, Fukushima Y, Liu C Yang S (2007) A gauge-based analysis of daily precipitation over East Asia. J Hydrometeor 8:607–626
60. Yatagai A, Arakawa O, Kamiguchi K, Kawamoto H, Nodzu MI, Hamada A (2009) A 44-year daily gridded precipitation dataset for Asia based on a dense network of rain gauges. SOLA 5:137–140. doi:10.2151/sola.2009-035

Remote Sensing Applications to Ocean and Human Health

Frank E. Muller-Karger
College of Marine Science, University of South Florida,
St. Petersburg, FL, USA

Article Outline

Glossary
Definition of the Subject and Its Importance
Introduction
Future Directions
Bibliography

Glossary

Attenuation depth Attenuation depth is a measure of how far electromagnetic radiation including light can penetrate into a substance. It is the depth at which the intensity of the radiation falls to 1/e (∼37%) of its original value immediately below the surface.

Diffuse attenuation coefficient The irradiance at a wavelength λ propagates over a distance (z) as determined by the diffuse attenuation coefficient. In aquatic environments, the diffuse attenuation coefficient is an indicator of the turbidity of the water.

Hyperspectral Hyperspectral data are collected by instruments called imaging spectrometers. These sensors are able to collect information from across the electromagnetic spectrum at a fine resolution of bands as narrow as 0.001 or smaller μm over a wide wavelength range, typically at least 0.4–2.4 μm.

Irradiance Irradiance is a radiometry term for the power of electromagnetic radiation per unit area at a surface. Irradiance is used when the electromagnetic radiation is incident on the surface, and it has units of watts per square meter (W/m^2).

Ocean color Ocean color is a general term used in the study of the biological and biogeochemical properties of ocean waters through remote sensing of the reflected and transmitted visible radiation. The "color" of the ocean comes from the interaction between light, water, and substances in the water, particularly phytoplankton (microscopic, free-floating photosynthetic organisms), detritus and inorganic particulates, and colored dissolved matter.

Radiance Radiance is a radiometric measure that describes the amount of light that passes through or is emitted from a particular surface area, contained within a given solid angle in a specified direction. It is used to characterize both emission and reflection from surfaces. The SI unit of radiance is watts per steradian per square meter ($W{\cdot}sr^{-1}{\cdot}m^{-2}$).

Thermal infrared radiation (TIR) Thermal infrared radiation refers to electromagnetic waves with a wavelength of between 3.5 and 20 μm. These waves are used to estimate the temperature of the surface of targets. This is a radiation typically emitted by objects as opposed to visible and short-wave infrared radiation which is part of the spectrum of sunlight reflected by objects.

Turbidity Turbidity is the relative clarity of a liquid and is an expression of the optical properties of water that causes light to be scattered and absorbed by particles and molecules rather than transmitted in a straight line through a water sample. It is a function of the concentration of suspended matter or impurities that interfere with the clarity of the water. Turbidity is a common index of water quality.

Visible, near infrared, and short-wave infrared (VIS, NIR, and SWIR) This broad band of electromagnetic radiation is used in remote sensing of the reflectance of the Earth. Light that is visible to the human eye is visible radiation (VIS) and encompasses a wavelength range from about 380–400 nm to about 760–780 nm. Near-infrared (NIR) radiation encompasses 0.75–1.4 μm in wavelength. The short-wave infrared (SWIR) is the wavelength range 1.4–3 μm. Together, this region of the spectrum is sometimes known as VSWIR.

Definition of the Subject and Its Importance

Remote sensing is defined here as the acquisition of information about an object without physical contact by way of recording or sensing devices mounted on aircraft, satellites, or simply sited on a high hill or bluff overlooking an area of interest. Ocean and human health is the general field that assesses conditions in the marine environment including estuaries that are relevant to the well-being of living resources and to the use of these resources and seawater by humans for amenities or the sustenance of life.

Remote sensing of environmental conditions in coastal and marine waters using space-based sensors has made great progress since the deployment of the Coastal Zone Color Scanner on the Nimbus 7 satellite in 1978. Remote sensing is an important complement to observations collected from field programs, because satellites provide estimates of a number of environmental parameters over large geographic areas (hundreds of square kilometers to globally) rapidly (in minutes), frequently, and periodically (often near-daily or better), and over the long term (in some cases, now upward of two decades). These observations support research, monitoring, and public outreach, providing a cost-effective complement to traditional monitoring efforts. Yet much work remains to be done to refine the technology to address the growing challenges to the health of marine ecosystems, including humans living within them, due to marine pollution, urban sprawl, overuse of resources, and changes in climate. This entry very briefly outlines the technologies used to conduct water quality assessments, including phytoplankton standing stock, turbidity, suspended sediment load, colored dissolved organic material (CDOM), diffuse light attenuation coefficients, temperature, salinity, wind stress, wave direction, amplitude and wavelength, and current speed and direction. The student and reader are encouraged to look for further detail on how to use these observations in the scientific and technical literature. A large number of international Earth-observing satellite missions are planned for the 2010–2025 time frame to monitor the marine environment. As part of this effort, a solid scientific base for remote sensing methods of marine pollution needs to be established, and multidisciplinary, international training programs need to be developed so that operational agencies can make effective use of these technologies. It is capable human resources that we are currently lacking the most.

Introduction

Satellite remote sensing of the Earth's environment has advanced significantly every decade since the late 1950s, when the first reconnaissance satellites were launched. There continue to be leaps toward more powerful personal computing, faster and more accessible global communication networks and media, standardization of a number of file formats, large and open-access databases, widely available photo-processing software, and online mapping and data visualization software including geographical information systems. These advances, combined with the growing computer literacy of broader segments of the world's population, permit the regular observation and application of several key terrestrial, atmospheric, and oceanic variables to understand how the environment changes at regional to global scales. Satellite-based sensors now routinely provide data that help interpret local environmental measurements in the context of large spatial scales, longer time frames, and of complex interactions between different processes. Satellite data facilitate the study of connectivity between ecosystems and human communities, of change against baseline environmental conditions over both small and larger geographic areas, and of possible natural or human-caused drivers of point and nonpoint sources of pollution and other threatening conditions.

This entry gives a brief overview of remote sensing applications in marine pollution, focusing on satellite instrumentation of singular interest, recent advances, and expectations for future directions. It is an update to the summary presented by Muller-Karger [1]. Only a few aspects of remote sensing can be presented here. For more detail on the history of the technology, the reader is referred to the references provided in that publication. While atmospheric pollutants are also relevant, here the focus is on detection of water properties or constituents in coastal regions.

Many relevant research efforts that provide a framework for analysis of the ocean's role in human health are coordinated by the International Geosphere-Biosphere Program (IGBP) and the World Climate

Research Program (WCRP). The Group on Earth Observations (GEO) has a 10-year plan (2005–2015) to implement a Global Earth Observation System of Systems (GEOSS) to connect remote sensing and users, fostering a global infrastructure to assess environmental change and its impacts on ecosystems including humans. A number of countries maintain space-based programs to map Earth resources and assess changes in the environment (Argentina, Brazil, China, several European countries as well as the European community, India, Israel, Japan, Russia, South Korea, the United States, and others). The Committee on Earth Observation Satellites (CEOS), which coordinates civil spaceborne observations of the Earth provides a useful online database on relevant missions, instruments, and measurements (http://database.eohandbook.com/). This database includes tables detailing mission dates, data characteristics, and applications.

Several national and international programs are also developing interfaces to facilitate merging of satellite data with other information to assist resource managers, policymakers, educators, and scientists. For example, NASA's SERVIR initiative integrates satellite observations, ground-based data, and forecast models to monitor and forecast environmental changes and to improve response to natural disasters. In 2006, the United Nations General Assembly agreed to establish the "United Nations Platform for Space-based Information for Disaster Management and Emergency Response Space-based Information for Disaster Management and Emergency Response" (UN-SPIDER), which seeks to further broaden access to space-based information to support disaster mitigation and response. Several other programs are being developed around the world, both by focused research efforts in academic institutions as well as by national and international government agencies. The reader is encouraged to follow the development of these global networks.

Measurable Variables

Remote sensing of the Earth's properties is based on the measurement of electromagnetic radiation reflected or emitted by a target. In terms of human health, all ocean waters are ultimately relevant as trends in global and regional ocean productivity and temperature are more than simply of academic interest. These tendencies, which can be assessed from space, may indicate large-scale changes related to climate variability [2–4]. They affect the pH of the upper ocean [5], affect lower trophic levels [6], and have impacts on higher trophic levels including fish [7, 8].

Estuarine, coastal, and continental shelf waters are directly relevant to human health. This is where much of the effort is concentrated on collecting resources from the ocean or to input our discharges. These waters support commercial and recreational boating, tourism, fisheries, and other industries and activities, and they receive materials from land either through rivers that drain inland areas, effluents from coastal cities, and from atmospheric deposition. There is a need to understand processes that change any factor in these waters that affect these uses, over both short and long time frames, as well as over short to synoptic space scales, including climate-scale changes.

While there had been some important attempts to apply remote sensing to coastal waters in the 1970s and 1980s, most of the research on coastal and estuarine waters conducted to date has been done more recently in the 1990s and the first decade of the new millennium. This research has focused on refining quantitative estimates of sediment load, turbidity or water clarity, algal concentration as a measure of eutrophication, detection of hydrocarbon slicks, and measuring temperature, bathymetry, and types of benthos. Studies by Epstein [9] highlighted the importance of remote sensing in addressing large-scale factors, including climate change, that promote the spread of disease in the coastal ocean.

Yet progress in the application of remote sensing to marine pollution, spreading of disease, or assessing the linkages between human and ocean health has been slow because few environmental parameters lend themselves to remote sensing. Often, these parameters are not easily measurable at the appropriate space and time scales. Pollution and disease vectors are time and again introduced to the marine environment at relatively small scales, and they disperse following complicated pathways. Limitations exist in observing these patterns using available satellite sensors, and the electromagnetic signal of the variables of interest is often masked or contaminated by other environmental factors. Clearly, the potential benefits of remote sensing are large, considering the high costs of monitoring using only traditional methods, but much work is needed to refine the technology to address even basic marine pollution problems.

Progress in this field has been made through interdisciplinary studies of indicator parameters designed to detect pollution or conditions that promote pathogen growth or their vectors by proxy or through studies of dispersal processes. Such inferences are best made when measurements are based on physical principles and when users are aware of technological limitations. For example, present state-of-the-art technologies and methods still do not permit direct estimates of nutrient, heavy metal concentrations, or pH, despite some early reports suggesting that this is possible (cf. [10, 11], and numerous unpublished reports). In many cases, any correlation between remotely sensed parameters and these variables is site specific, ephemeral, and may be coincidental, necessitating statistical validation on a site-specific, image-per-image basis.

Phytoplankton standing stock, estimated in terms of chlorophyll or "pigment" concentration, is an important indicator of water quality. Among other relevant variables that may be sensed remotely are diffuse light attenuation coefficients, suspended sediment load, "turbidity," dissolved organic material (yellow substance), temperature, salinity, wind stress, wave direction, amplitude and wavelength, and current speed and direction. This entry provides a very brief overview of methods used to detect these variables.

Approaches

The most successful approach to remote sensing combines a variety of measurement platforms and data types, including automated in situ and shipboard observations for detailed ground truth and remote sensors to cover a range of spatial and temporal scales. Satellite and airborne remote sensing is typically limited to signals originating from the surface or "skin" of the ocean (microwave and infrared radiance) or, at most, a few attenuation depths from the surface of the sea (for visible light, the attenuation depth is the distance over which radiance decays by a factor of 1/e; c.f. [12]). It is therefore very important to obtain complementary measurements at greater depths to resolve one of the most important dimensions in the aquatic environments, namely the vertical. Similarly, since remote sensing only provides direct estimates of a few parameters, it is critical to establish the empirical or deterministic relationship with any proxy that may be used to asses ocean health.

Advances in telemetering of data from remote sampling sites and platforms (e.g., [13–15]) promise solutions to a range of environmental monitoring problems, including in situ measurement of chemical processes [16], contaminants [138], nutrient concentrations [17–19], precipitation [20, 21], bioluminescence [22–24], zooplankton and larger animals [25–27], light (c.f. [28–30]), and primary productivity, and other variables (see [31], and references therein). This is an area of active development that will provide important calibration and supporting data for remote sensing work, thus facilitating ground truth efforts.

The advantage of satellites is that they augment the scope of oceanographic studies by providing frequent and repetitive coverage of large spatial scales, and that the user is rarely involved in the complicated process of collecting the raw data [32]. Many of these synoptic time series now extend for many years, and decadal datasets are now available. This permits evaluation of anomalies relative to a specific time frame and comprehensive comparative studies. However, satellite data is voluminous, and processing is computationally intensive, which leads to trade-offs between spatial and temporal resolution. Spatial resolution is also sacrificed to improve radiometric sensitivity, which is critical because many signals are weak and because target identification is usually based on narrow features in reflectance spectra.

Aircraft, including helicopters and balloons, are platforms on which prototype satellite instruments are frequently tested. Instruments flown on aircraft provide finer spatial and temporal resolution than their satellite counterparts, flexible flight patterns and altitude, and short deployment times. These features frequently make aircraft the platform of choice during catastrophic pollution events (see, e.g., [33, 34]). However, aircraft fitted for remote sensing are expensive and often are not readily available, a situation that may be at least partially remedied by drone planes and helicopters. Also, while merging observations with navigation data has become easier, sunglint, surface wave patterns, and contamination by unmeasured factors (e.g., cloud shadows, electronic aircraft system noise, calibration drift, etc.) are common.

Applications and Sensors

Since the 1970s, there have been significant advances in the development of sensors flown on international satellite missions and on aircraft to measure coastal, estuarine, and marine parameters of interest to ocean and human health. Extremely sensitive sensors are required to measure the visible sunlight reflected by materials suspended in the surface layers of the ocean or lying on the bottom in coastal waters shallower than about 20 m. Sensors finely tuned to specific infrared wavebands to which the atmosphere is relatively transparent are used routinely by many sensors to measure the temperature of lakes, coastal waters, and the world's oceans several times per day, both during daytime and nighttime. A number of microwave sensors are used to estimate wind and surface current speed and direction, temperature, ice cover, and even salinity. The satellite sensors therefore afford rapid, repeated, and synoptic assessment of a number of environmental parameters concurrently anywhere on the planet, at pixel resolutions that typically range between a few meters to hundreds of kilometers and at particular times of the day, at a relatively low or no cost to the user. Airborne sensors provide much higher spatial resolution spanning centimeters to meters and the ability to sample a particular time of the day, but logistics are expensive, and obtaining good data quality in aquatic environments is challenging.

Remote sensing data have been tested in a number of applications over the years. Many attempts have been made to link various observable parameters to factors that affect ocean and human health. Much work remains to be done with both historical datasets as well as observations from new sensors and combining data from various sensors and automated or discrete ground observation platforms. Today, the most significant application in coastal and marine observation remains the use of various sensors to characterize a number of environmental parameters concurrently over synoptic scales rapidly and repeatedly and within the context of long time series. This allows detection of change and anomalous conditions relative to some previous condition or arbitrarily constructed "climatology." Some applications are extremely useful but still as of yet mostly underutilized by the public, resource management officials, and the public health community. Some other applications have great potential, but as of yet they have not been proven and have been oversold. A few of these applications are described in the next few paragraphs under the headings of specific technologies.

Ocean Color Among the most useful indicators of water quality is ocean color. The color of the ocean is the result of the interaction between visible solar irradiance, water, and substances like phytoplankton, suspended sediment, colored dissolved organic matter, or with the bottom of the ocean. The color contains much information about the condition and health of water bodies [141].

Ocean color studies fall broadly under the category of aquatic bio-optics, and their objective is to gain insight on processes such as the distribution and dispersal of dissolved organic carbon, changes in oceanic primary productivity, targeting and identifying coastal phytoplankton blooms, conducting regional and global biogeochemical assessments, understanding changes in benthic communities in clear, shallow tropical coral reefs, and monitoring coastal water quality. Sensors with high, medium, and low spatial resolution have been used to develop and test applications to assess various parameters relevant to ocean and human health. Most attempts at using high spatial resolution sensors have not led to accurate assessments nor developed into widely used operational tools. At this stage, the best promise for remote sensing of biogeochemical processes in the coastal ocean remains with sensors that have high sensitivity and high dynamic range and frequent coverage. This restricts the type of sensors to those having narrow spectral bands, medium spatial and spectral resolution, repeat orbits that provide daily or near-daily observations, and well-calibrated visible radiance data, i.e., those that are commonly known as "ocean color" instruments.

An important issue that requires significant attention is the estimation of the atmospheric radiance seen by an ocean color sensor [35, 36]. Over 80–90% of the total visible radiance received by a satellite sensor looking at the ocean is due to atmospheric scattering of solar radiation or specular reflection of light by the ocean's surface. These sources of radiance need to be estimated with high precision and accuracy exceeding 1% since an error of this order will lead errors that can exceed 10% in the water-leaving radiance. The water-leaving radiance is light backscattered by the water column, not by the surface, which is detected by a downward-looking sensor just above the surface. In addition, marine waters are dark compared to land, and coastal and estuarine waters are especially dark since materials suspended or dissolved in them typically absorb visible light strongly across the spectrum. Therefore, ocean color sensors have to be extremely well calibrated and have signal-to-noise ratios exceeding 500 to over 1,000 the typical radiance values observed over the ocean. Among the most used ocean color sensors are the following:

Satellite sensors (see CEOS; http://database.eohandbook.com/ for a complete list):

- Coastal Zone Color Scanner (CZCS; 1978–1986; NASA/USA).
- Sea-viewing Wide-Field-of-view Sensor (SeaWiFS; 1997–2010; NASA and Orbimage/GeoEye; USA).
- Moderate Resolution Imaging Spectrometer (MODIS on the NASA Terra and Aqua satellites; 1999 and 2000 launches and continuing in 2011).
- Medium Resolution Imaging Spectrometer (MERIS on the ESA ENVISAT satellite; 2002 launch and continuing in 2011).
- Ocean Color Monitor (OCM; on several ISRO Oceansat satellites; from 1999 and continuing in 2011).
- Visible/Infrared Imager/Radiometer Suite (VIIRS, planned for launch in late 2011 on the US National Polar-Orbiting Operational Environmental Satellite System Preparatory Mission/NPP and on the future Joint Polar Satellite System/JPSS with a planned launch date of 2015).
- Hyperion (on the NASA EO-1 satellite; launched in 2000 and continuing in 2011).
- Several other international missions.

Aircraft sensors:

- Airborne Visible/Infrared Imaging Spectrometer (AVIRIS).
- Compact Airborne Spectrographic Imager (CASI).
- Airborne Oceanographic Lidar (AOL).
- Several other research and commercial sensors.

Applications of Ocean Color Data Ocean color data permit a wide range of observations relevant to ocean and human health applications. Ocean color observations are of particular interest because of their ultimate utility in estimating the spatial distribution of stocks and rate of change of various organic carbon pools in the ocean, including those associated with living and detrital particulate materials and with colored dissolved organic matter. In addition, these data are useful for coastal applications including monitoring of suspended particulate and dissolved materials associated with river discharge and resuspension from the bottom and, in general, for assessing coastal water quality. This is accomplished primarily by retrieving chlorophyll concentration and absorption and backscattering properties of the water from spectral measurements of the surface reflectance of the ocean.

Attempts to quantify chlorophyll concentration, colored dissolved organic matter (CDOM) absorption, suspended sediment load, "turbidity," and the diffuse attenuation coefficient of light as an index of water clarity have been made using a variety of satellite sensors. Geographically, of particular interest are coastal and estuarine waters, where so many human activities are focused. River plumes often lack an infrared (IR) signature that distinguishes them from marine waters, especially as distance from the estuary increases. However, plumes are strongly colored relative to sea water. Ocean color sensors are characterized by their narrow spectral bands, very sensitive sensors, and spatial resolution of between ca. 250 × 250 m^2 and 1 × 1 km^2 pixels. The advantage of these sensors is that they cover areas $>10^6$ km^2 in only 2–3 min, with revisit times of 1–3 days.

The CZCS [37] proved the value of using ocean color satellite images for tracing plumes over long distances. CZCS bands were centered at 443, 520, 550, 670, 750, and 11,500 nm and sought to match chlorophyll absorption maxima and minima. The red channels allowed atmospheric correction. Since the CZCS was an experimental sensor, data coverage was not continuous, the instrument was less sensitive than required for coastal observations, and the calibration was hard to establish. In effect, the technology was extremely useful to visualize, for the first time, global ocean phytoplankton biomass and river plume dispersal in the open ocean; the CZCS did not allow accurate estimates of coastal water colors, coastal chlorophyll concentrations, or turbidity. The CZCS ceased operating in 1986, leaving behind a very large data set [38].

Among the most relevant ocean and human health studies were the pioneering studies of Barale et al. [39]. They used the CZCS to help calibrate numerical simulations of the circulation of the Adriatic Sea and the Po River plume to provide information on seasonal dispersal of contaminants in the region [40, 41]. Maynard et al. (1987) then used the CZCS to map the dispersal of Yukon River water and that of smaller rivers in Norton Sound (Alaska) to determine the area of impact of potential pollutants on the northern Bering Sea. Holligan et al. [42] compiled an atlas of CZCS and AVHRR images to document advection and mixing processes in the North Sea, where human impact is intense.

In an attempt to outline seasonal circulation patterns of the tropical western Atlantic, the Caribbean Sea, and the Gulf of Mexico, and to evaluate the effect of nutrient inputs via rivers into these oligotrophic seas, Muller-Karger et al. [43, 44] illustrated the seasonal dispersal of the Amazon, Orinoco, and Mississippi river plumes. CZCS imagery showed the marked seasonal variability of these plumes. These papers illustrated the connectivity between distant locations in ocean basins, with river plumes often extending hundreds to over 1,000 km from their source. Müller-Karger et al. [45] also combined ocean color and infrared imagery (from the AVHRR) to trace eddies and other oceanic circulation features in the Gulf of Mexico. Müller-Karger et al. [45] and del Castillo et al. [46] used CZCS and SeaWiFS images to demonstrate that the Mississippi plume does not always disperse to the west of the Mississippi Delta but that it also disperses to the east and south in the Gulf of Mexico and that it can reach the Florida Keys and disperse along the eastern seaboard of the USA. The significant river influence over shelf areas can be appreciated in satellite images of many river deltas and river plumes. These studies laid some of the groundwork for analyses of potential dispersion pathways of contaminants like oil.

There effectively was a 10-year gap between the end of the CZCS mission in 1986 and the resumption of ocean color observations with the Japanese OCTS on the ADEOS-I mission. The field of ocean color research and applications moved forward significantly with the launch of the SeaWiFS sensor in 1997 shortly after the failure of the ADEOS-I satellite, and the MODIS and MERIS sensors in 1999 and 2000 as the twenty-first century began. Oceanic remote sensing moved toward complex and multidisciplinary studies of environmental change over a wide range of time and space scales. These studies merged observations from various satellites. A concerted effort was initiated to understand coastal and turbid waters [47].

The differences between sensors and missions are considerable, ranging from orbit and ground revisit time to swath width, spectral and spatial resolution, sensitivity, accuracy, and so on. Examining each is beyond the scope of this review. For the purposes of illustrating applications, NASA's Moderate Resolution Imaging Spectroradiometer (MODIS, [139]) and Orbimage's SeaWiFS [144] are focused on because of their multiyear time series, global coverage, and their high-quality observations. The SeaWiFS sensor ceased operations in 2010. MODIS sensors continue data collections, and all data from the CZCS, SeaWiFS, and MODIS satellite sensors are public. The MODIS model is of particular interest because it is the prototype for the Visible/Infrared Imager/Radiometer Suite (VIIRS), which will fly on the US National Polar-Orbiting Operational Environmental Satellite System Preparatory Mission (NPP) and the future Joint Polar Satellite System (JPSS).

The overall accuracy of retrieved geophysical parameters (water leaving radiance, CDOM absorption, chlorophyll concentration, etc.) depends on the performance of the atmospheric correction and in-water algorithms. Of primary concern are sensor calibration and characterization. Errors in water-leaving radiance estimates can easily by exacerbated or exaggerated in coastal and estuarine waters because the blue water-leaving radiance is often much smaller than that in clear ocean waters. Significant research has focused on improving the traditional multiband chlorophyll retrievals. The MODIS and MERIS sensors are also equipped with spectral bands specifically designed for measuring chlorophyll fluorescence relative to a baseline (chlorophyll fluorescence line height algorithm; [48]). The chlorophyll fluorescence efficiency varies in time and space, which is an area of research. Aggregating such data with those from other bio-optical algorithms helps address high-chlorophyll coastal waters (e.g., [49–51]). Unfortunately, the VIIRS sensor lacks the capability to detect this solar-stimulated fluorescence, limiting its capability for coastal observations relative to MODIS and MERIS.

Water constituents are estimates based on the water-leaving radiance. Algorithms related to biological oceanographic properties are referred to as "bio-optical" algorithms. For Case I waters, where one variable (phytoplankton) dominates color, a band-ratio (blue/green) bio-optical algorithm is often effective to estimate chlorophyll-a concentration because phytoplankton absorbs more blue light than green. Robust band-ratio algorithms have been used for CZCS [52], SeaWiFS, and MODIS [53]. In turbid coastal waters where constituents do not co-vary and phytoplankton does not dominate water color, band-ratio algorithms often fail (e.g., [50]). Semi-analytical algorithms may help in these cases [54–57]. The objective is to differentiate between various constituents, namely chlorophyll, colored dissolved organic matter (CDOM), and total suspended solids (TSS). While some progress has been made over time (e.g., [50]), assumptions need to be fine-tuned regionally, as Case II waters vary in space and time. Examples of parameters that are difficult to estimate but that are required in these optical models are the spectral slope of CDOM absorption, particle backscattering, bottom reflectance, and depth.

There has been significant progress in estimating and mapping the distribution of colored dissolved organic matter in the oceans (see [58–60], and references therein) and in coastal waters. The work of the SWFDOG group [49] illustrates the importance of using time series of satellite images to detect and track anomalies in coastal water quality. They detected an event off western Florida, USA, which the media called a "black water" event and which caused significant anxiety among local coastal residents, divers, and fishermen. Daily SeaWiFS data collected between September 1997 and August 2002 showed this unique event that lasted over 4 months. The "black water" was advected into the Florida Keys National Marine Sanctuary, where it led to the death of corals and sponges. By tracking the event, it was possible to determine that the extensive dark water patch evolved from a senescent bloom that was stimulated into activity again by local river input. The dark color of the water was due to high concentrations of CDOM and decreased backscattering. A field survey showed chlorophyll concentration was 5–10 mg m^{-3} due to a nontoxic diatom (Rhizosolenia) bloom.

While chlorophyll is a parameter desired by many coastal managers, it remains very difficult to estimate with sufficient accuracy in coastal and estuarine waters. A small error in CDOM assessment significantly affects blue absorption estimates and causes large errors in chlorophyll retrievals. For example, the CDOM absorption coefficient at 443 nm in an estuarine area may be $0.3-0.4\ m^{-1}$ or larger, while a chlorophyll concentration ~1 mg m^{-3} effects an absorption coefficient of 0.03–0.04 m^{-1} at 443 nm. Thus, a 10% error in CDOM estimates leads to 100% errors in chlorophyll estimates. Throughout the "black water" and other similar events off western Florida, it was possible to differentiate between coastal plumes that contained a phytoplankton bloom and those that were dominated by CDOM by examining both the blue and green water-leaving radiance bands of SeaWiFS and MODIS and the fluorescence line height (FLH) images from MODIS [50].

Satellite ocean color sensors have shown great advantage over traditional means to monitor the spatial extent and duration of harmful algal blooms (HAB) or red tides. The state-of-the-art primary indicator of a bloom (toxic or not) is high chlorophyll concentration. At present, there is still no reliable algorithm to discriminate between red tides and other blooms, such as one dominated by diatoms or by other nontoxic phytoplankton. For example, during Alexandrium blooms which occur in the Gulf of Maine off the northeastern United States, these organisms co-occur with significant quantities of other phytoplankton, and therefore, waters do not seem to have a particular spectral signature that would allow identification of this HAB.

Chlorophyll concentration has been used as a proxy to build a crude operational monitoring tool for Karenia blooms off Florida (e.g., [61]). However, reliable estimates of chlorophyll concentration in these coastal waters have not been achieved, and bio-optical algorithms are still the focus of intense research [62–64]. A fundamental difficulty is the presence of CDOM since a small error in CDOM estimates generates large errors in computed chlorophyll values. A positive chlorophyll concentration anomaly also does not represent proof positive of a red tide bloom. Therefore, false positives are inevitable, particularly during periods of higher river discharge or when other blooms occur. At this stage, the capability to detect an anomalous chlorophyll concentration provides an important advantage in planning proper field surveys and other responses by resource managers.

Great expectations revolve around identification of phytoplankton species by spectral analyses combined with other environmental observations [65, 140]. A recent study, for example, found that the backscattering to chlorophyll ratio is generally lower in *Karenia brevis* blooms relative to other blooms [66]. The proposed explanation is that this species experiences less grazing. How to take this scientific discovery to an algorithm and into an operational system is being investigated.

A useful and important application of remote sensing data is the assessment of suspended solids in aquatic environments. High concentration of suspended matter is an indication of erosion, resuspension of benthic sediments that may have contaminants, or of eutrophication due to a variety of causes. This condition also blocks sunlight from reaching benthic algae and sea grass, negatively affecting the health of coastal and estuarine waters. Turbidity is an index of light attenuation and water quality commonly used in estuarine and coastal areas. Several studies have demonstrated the utility of remote sensing to estimate turbidity in coastal, turbid waters. Miller and McKee [67] used MODIS 250 m data to assess total suspended matter in Lake Ponchartrain and adjacent waters in Mississippi, USA. Chen et al. [68, 69] and Moreno et al. [70] examined time series of 250 m MODIS images of Tampa Bay and showed that the synoptic satellite observations, collected every 1–3 days, are an important complement to the monthly water quality sampling program carried out in the bay. They illustrated rapid changes in spatial and temporal water clarity and turbidity conditions that were missed by the in situ surveys and were able to explain turbidity changes related to tidal, wind-driven, and discharge events. Rodríguez-Guzmán and Gilbes-Santaella [145] used 250 m MODIS data to estimate suspended matter concentrations in Mayaguez Bay, Puerto Rico. Similar studies have been conducted in France by Doxaran et al. [71].

An interesting conceptual framework for coastal ocean and human health applications was defined by Lobitz et al. [72] and Colwell [73], who identified a need for synoptic observations of phytoplankton blooms and sea surface temperature to help assess the potential for cholera outbreak conditions in coastal and estuarine waters. Colwell [73] and Hu et al. [74] further made the case that climate change affects the distribution and frequency of diseases, many of which spread along coastal zones with biological carriers. Colwell [73] found that the spread of the phytoplankton blooms and associated cholera outbreaks throughout the tropical and subtropical Pacific Ocean coasts of South America were associated with El Niño-Southern oscillation events. Hu et al. [74] found a linkage between climate variability, as quantified by the Southern Oscillation Index (SOI), and dengue fever epidemics in Queensland, Australia. They found an increase in the numbers of dengue fever cases 3–12 months after a decrease in the average SOI (i.e., warmer conditions).

Several attempts have been made to use the Advanced Very High Resolution Radiometer (AVHRR) on the NOAA Polar Orbiters to estimate coastal and oceanic water quality parameters [75–77, 146, 147]. The AVHRR (see [78]) has a nominal pixel resolution of 1 × 1 km and near-infrared bands that are sensitive to reflected sunlight and which have a dynamic range that accommodates the high reflectance of land and coastal waters. Since it is an operational facility, data are available several times a day, worldwide. Stumpf et al. found that when calibrated with concurrent in situ observations, individual AVHRR images provided information on sediment load and phytoplankton concentration in estuaries. They point out that the technique can be improved by correcting for atmospheric effects.

While the moderate-resolution class of ocean color and near-infrared sensors (historically, the CZCS, SeaWiFS, MODIS and MERIS, and AVHRR) allow observation of large-scale phenomena, there is a need for high spatial resolution data to address local marine pollution problems. Landsat data from the Thematic Mapper class of sensors are available for many areas around the globe since the mid-1980s and are now free to the public and distributed by the United States Geological Survey (USGS). Attempts to map coastal and oceanic water quality parameters have been made since the 1970s, initially using Landsat sensors ([79–85]; and others). Use of Landsat and the French SPOT sensors has grown rapidly because they provide pixel resolutions of 30–80 m in spectral mode. They have been used to map suspended sediment and variations in water color in the nearshore environment [81, 84, 86–89, 143].

Since the late 1990s, higher spatial resolution imagery from several commercial satellites has become available, although at considerable cost to the user. These data are used extensively to map general land features, for urban and land-use planning, resource assessment, and disaster response efforts. Hellweger et al. [90] provide an example of the utility of high-resolution (1 m IKONOS) multispectral satellite imagery for estimating water quality patterns in the lower Charles River (Boston, USA), with limited results. Single image applications do not satisfy coastal and ocean health studies or applications. These important applications will continue to be limited as long as the data remain costly because building any time series and/or climatology against which to assess change is prohibitively expensive.

However, compared to the ocean color sensors mentioned above, past Landsat-class sensors have broad bands; for example, the "blue" channel on Landsat TM and ETM + (0.45–0.52 μm) spans the blue chlorophyll absorption peak and blue-green shoulder, the "green" band (0.52–0.60 μm) spans part of the blue-green shoulder, the green absorption minimum, and the green-yellow shoulder. SPOT lacks blue bands. The Landsat Data Continuity Mission (LDCM), planned for launch in late 2012 or 2013, includes a new Coastal/Aerosol band spanning 0.433–0.453 μm. This will be a very important tool in coastal resource assessment.

Most attempts to evaluate concentrations of in-water constituents with Landsat or SPOT have been based on scene-specific, concurrent in situ observations to provide a statistical base. In general, good qualitative information may be obtained in the immediate nearshore environment (e.g., [91]). At best, they may provide similar information to that derived from the AVHRR in highly turbid environments using red and near-IR bands but with much better spatial resolution [92]. The high cost of these data and the long revisit times (8–17 days due to orbital characteristics and narrow swaths) precludes detailed time series analyses or even use as monitoring tools.

Two significant advantages of Landsat-class data are that historical data are available since the mid-1980s and this allows comparison of a number of parameters including shoreline location and that Landsat data are now free of cost. However, Landsat and most other high spatial resolution sensors data provide only a limited capability to conduct systematic studies of waterborne constituents such as phytoplankton or sediment concentration either in coastal zones or elsewhere because of the limited radiometric, geometric, and revisit characteristics of the sensors. They are not well suited for quantification of aquatic chlorophyll in open waters and much less in areas of high chlorophyll because of the problems listed above. An important issue is the very high-cost per square kilometer of data for data from most of the commercial high-resolution sensors. These data are of great interest to the science community, but research using these data has been hampered by their cost.

One of the most exciting developments of the last decade is the availability of data from a series of proof-of concept hyperspectral satellite sensors. Hyperspectral data provide a significant improvement over the traditional multispectral sensors that provide between four and ten bands in the visible to differentiate between various optically active constituents present in coastal and estuarine waters [93]. The most widely accessible sensor is the NASA Hyperion sensor, launched in 2000 aboard the EO-1 proof-of-technology satellite. Hyperion collects images in 220 spectral bands at 30 m resolution, compared to the ten multispectral bands flown on traditional Landsat missions. The instrument collects images spanning 7.5 km by 100 km. The Hyperion data are new to coastal water quality assessments. Brando and Dekker [94] tested the Hyperion data and found that the sensor sensitivity is sufficient to estimate colored dissolved organic matter, chlorophyll, and suspended matter in Moreton Bay, an estuary in Australia. The Hyperspectral Imager for the Coastal Ocean (HREP-HICO; [95]), installed aboard the International Space Station since 2009, operates a visible and near-infrared (VNIR) Maritime

Hyperspectral Imaging (MHSI) system. The goal is to detect, identify, and quantify coastal geophysical features and to test algorithms for water clarity, chlorophyll content, water depth, and ocean or sea floor composition for civilian and naval purposes.

Aircraft Sensors Aircraft instruments play an important role in monitoring aquatic pollution at small scales. The AVIRIS, flown on NASA's ER-2 high-altitude aircraft (a U2), provides 224 bands of visible/infrared data at ~20 m resolution [96–98]. Its sensitivity for detecting variations in water-leaving radiance is relatively low but adequate where highly reflective constituents occur. Karaska et al. [99] used AVIRIS images of the Neuse estuary in North Carolina to estimate the concentrations of chlorophyll, suspended matter, CDOM, and turbidity and created an index of eutrophication for these waters. Bagheri and Yu [100] applied similar techniques to map water quality in the Hudson/Raritan estuary.

Another aircraft instrument is the FLI, also called the Programmable Multispectral Imager (PMI), of Moniteq Inc., Canada [101]. Dekker et al. [102] examined variations in eutrophication of the Loosdrecht Lakes in the Netherlands with the PMI and concluded that this instrument was ideally suited to examine distribution of pigments because of its high spatial resolution and sensitivity at long wavelength bands (ratios of bands between 600 and 720 nm were needed).

Among popular aircraft instruments used for water quality monitoring is the CASI of Moniteq Inc., Canada [103, 104]. Unlike others, this instrument is well calibrated, small (about the size of a personal computer), and may be used in an "imaging" mode with a few selectable bands or in a "spectral" mode in which a few pixels are collected for 288 bands between 450 and 950 nm for limited "look directions." The applicability of the CASI to marine pollution problems still needs to be better documented.

Active sensors, namely those which illuminate a target and measure electromagnetic radiation returned to the system, also show promise in water quality studies [105]. Among the most prominent is the NASA AOL. The AOL can be operated in a LIDAR mode (time gated) or as a multispectral analyzer with or without a fixed time (depth) delay [32]. This laser sensor presents many advantages, among which are specificity of response, relative simplicity of signal interpretation, low-altitude operation (including below cloud ceiling), and the potential of providing depth distribution of relevant variables by time-gating laser pulses [32]. Hoge and Swift [106] demonstrate the high correlation between fluorescent return and chlorophyll. This instrument is also useful for mapping horizontal and vertical turbidity variations [107].

Other Measurable Variables

Oil Spills, Surface Organic Slicks, Surfactants, and Yellow Substance Traditional remote sensing techniques to address oil spills in aquatic environments include optical (passive visible and infrared, laser fluorosensors), and passive and active microwave (e.g., Synthetic Aperture Radar, SAR) using aircraft or satellites (see [108, 109, 148, 149]). Most of the satellite SAR sensors and the various airborne sensors used to detect oil at the surface of the ocean are expensive and do not provide the required high-frequency coverage. The MODIS sensors include bands that generate images at 250 and 500 m spatial resolution. These have great unexploited potential for coastal monitoring because they can be used to study small-scale features. One application is in assessments of oil spills and water quality where 1-km satellite data are often inadequate. One example was illustrated by Hu et al. [51]. The medium-resolution (250- and 500-m) MODIS Level-1 total radiance imagery from 1 December 2002 to 9 March 2003 was examined, and patterns were found within Maracaibo Lake that is suspected to be extensive spill patches; the presence of oil was confirmed by the ground surveys.

This experience was important as it helped to rapidly implement the use of MODIS and MERIS data to map surface oil features during the British Petroleum Deepwater Horizon (BP DWH) disaster in 2010. Satellite remote sensing was essential in mapping and tracking surface oil dispersal during this event (Muller-Karger, 20 May 2010, testimony before the House Subcommittee on Energy and Environment, US Congress). Exploring the promise of this technology is important to understand the potential of the VIIRS imaging bands on the NPP and JPSS satellite platforms.

Organic films on the ocean surface can be detected and mapped by laser light [110]. In 1978–1979, a series of oil spills were made off Sandy Hook, New Jersey, in waters >40 m depth. These spills were treated with helicopter-deployed dispersants containing surfactants soon after deployment and overflights with the NASA AOL were used to map the spills [111]. The results show that Raman backscatter strength can be related to oil-film thickness after removal of background and oil fluorescence contributions [111, 112]. Laser light can also be used to determine the spatial distribution of dissolved organic material in marine and estuarine environments [113, 114].

Infrared imagery such as that of the AVHRR has also proved useful to trace very large oil spills. During the Persian Gulf War of January–March 1991, oil slicks could be traced off the coast of Kuwait as sea surface temperature (SST) anomalies in the nighttime imagery. Such patterns were undetectable in the daytime AVHRR imagery (O. Brown, U. Miami, personal communication 1991).

Oil contamination of the Persian Gulf could also be traced in handheld photography collected by Space Shuttle astronauts [115]. During the Iran-Iraq war, oil slicks emanating from burning oil tankers and drilling platforms along the Iranian coast could be seen in the photography. Similarly, oil released from Kuwaiti loading facilities and oil fields during the 1991 Gulf war could be seen spreading along the Kuwait and Saudi Arabia shorelines in sunglint photography.

Microwave sensors also have an application in monitoring oil pollution in the ocean. In particular, the SURSAT (Surveillance Satellite) program, implemented in 1977 by the Canadian government, proved the utility of microwave sensors in monitoring ocean traffic, ice coverage, weather, and ocean pollution [116]. SURSAT provided the basis for participation by Canada in the US SEASAT program. The NASA SEASAT satellite, which operated between 26 June 1978 and 10 October 1978, was a test-bed for a variety of ocean-looking instrumentation. Among the most successful instruments was a Synthetic Aperture Radar (SAR), with 25 m resolution, which seemed able to detect oil spills. However, the subtle distinctions between oil slicks and natural phenomena resembling oil slicks require further work [116, 117].

Temperature Satellite instruments designed to measure sea surface temperature (SST) are among the most successful for both research and operational applications. Several instruments now provide routine, daily coverage of the world's coastal and marine environments, including the NOAA and MetOP AVHRR, and NASA MODIS. They provide observations that facilitate tracing of circulation features of spatial scales ranging from 10 to 1,000 km (c.f. [45]). These sensors are also used to track weather patterns and map land vegetation indices [118, 119].

An important application of satellite SST is detection of temperature extremes in coastal environments [120, 121]. Temperature variability can both enhance the resilience of coastal and marine organisms, but both cold and warm extremes can be fatal. Soto [122], for example, documented larger benthic coral reef cover in coastal areas of the Florida Keys, USA, where SST variability was larger than where SST was more stable. Satellite images are now also routinely used to assess areas around the world at risk of coral reef stress due to high-temperature anomalies [123].

Temperature contamination frequently occurs at smaller scales (<<100 km) as a result of industrial discharge, and the sensors with 1 km spatial resolution such as the AVHRR and MODIS are generally inadequate to resolve these scales. The Landsat TM and ETM + sensors provide a useful tool to assess surface temperature of aquatic environments at such scales, with Thermal Infrared (TIR) data at 60 m pixel resolution. For example, Thomas et al. [124] and Fisher and Mustard [125] used Landsat data to study sea surface temperature patterns off the coast of New England, USA. These studies demonstrated that it is possible to assess the dominant seasonal patterns in cross-shelf SST gradients at scales of small embayments and estuaries in coastal zones around the world. A limitation remains the repeat cycle of the Landsat-class satellites, i.e., every 16 days.

An instrument that provides higher resolution is the Precision Radiometric Thermometer Model 5 (PRT-5), deployed from aircraft or even ships. This sensor provides an operational capability to map along-track SST.

Salinity Remote sensing of sea surface salinity from aircraft and satellites, using microwave radiometers tuned to frequencies of order of 1 GHz, has been limited primarily by the poor spatial resolutions obtained with conventional microwave antennas (i.e., >500 km from satellites). Proven aircraft capabilities are accuracies <10 practical salinity units and 0.2 units in resolution. Lagerloef et al. [126] discuss synthetic aperture antenna technology that would allow resolutions of ca. 10 km from space and higher resolutions from aircraft altitude.

Two missions are now poised to provide global satellite observations of sea surface salinity. The European Soil Moisture and Ocean Salinity Satellite (SMOS), launched in November 2009, is part of ESA's Living Planet Programme and was designed to facilitate the study of the Earth's water cycle and climate. The Aquarius/SAC-D mission, launched in June 2011, is a focused satellite mission to measure global sea surface salinity. The Aquarius/SAC-D mission was developed by NASA and the Space Agency of Argentina (Comisión Nacional de Actividades Espaciales, CONAE). These instruments are expected to have accuracies of 0.5–1.0 psu and be unaffected by cloud cover. The limited spatial resolution and accuracy, however, will be of limited use in pollution monitoring work.

Ocean Currents, Wave Height and Direction An important capability afforded by satellites is the ability to map sea surface topography (sea level variations in space), sea level changes, and the roughness of the ocean's surface over large to global scales, repeatedly and over long periods of time. This is possible with a number of active microwave sensors (radar). Most of the radar sensors that fall under the category of "altimeters" and "scatterometers" have relatively low spatial resolution (tens of kilometers), which limits direct application to examine physical oceanographic conditions within estuaries and most coastlines. Synthetic aperture radars (SAR) on the other hand have very high spatial resolution (order of meters) and are helpful to map larger waves, including internal waves. SAR data are also very useful to map characteristics of the texture of the sea surface that may be a telltale of pollution or other conditions relevant to ocean and human health. As mentioned above, SAR data have been instrumental in mapping surface slicks, oil spills, and natural oil seeps in the ocean.

Altimeter radar sensors can today accurately measure changes in sea surface topography in the order of 1–2 cm over areas spanning from several tens of square kilometers to global. Among the most important applications of altimeters are measuring sea level and the variability in sea surface height in an ocean basin ([127, 128, 142]). The importance of understanding regional and global trends in sea level over short and longer periods cannot be overemphasized. Altimeter observations show that sea level around the turn of the millennium rose at an accelerated rate of about 3.1 mm per year. This is significantly higher than the average rate for the twentieth century and is now considered a major threat to coastal communities by the Intergovernmental Panel on Climate Change (IPCC, [129]). Sea surface height variability also provides a means to measure regional circulation in the interior of the ocean over large scales, but these measurements have very large uncertainties near the coast and over continental shelves because of limitations in the technology and in the ability that the altimeter data has to be corrected for gravitational factors. A summary of coastal applications may be found in Emery et al. [130].

Radar altimeters are also useful to help characterize waves in the ocean [131–133]. Knowledge of ocean currents and waves is important to ocean and human health in a number of ways. For example, they affect ship routing in the ocean, they are relevant in search and rescue efforts, they are useful to fishermen or recreational boaters, and they help understand coastal erosion and flooding patterns.

Wind Speed Wind is relevant to marine pollution because of its role in the generation of currents, waves, and the direct dispersal of contaminants. Passive remote sensing of wind is possible using visible, infrared, or microwave imagery of the ocean. Using series of visible and infrared images, wind fields may be derived by tracking identifiable clouds. However, extrapolation of such winds to the surface and lack of coverage can render these techniques unsatisfactory.

Instead, microwave emissions by the sea surface have been empirically related to surface wind speed, and several experiments were conducted with Scanning Multichannel Microwave Radiometers (SMMR) carried aboard the SEASAT and Nimbus-7 satellites to refine this relationship. The understanding of the relationship between emissivity of the ocean and wind characteristics has grown substantially since the late 1970s. Unfortunately, large uncertainties in the results occur due to sensor calibration drifts. Somewhat better accuracies are possible with the Special Sensor Microwave/Imager (SSM/I), presently operational on the Defense Meteorological Satellite Program (DMSP) satellites. Swift [134] summarizes the physical basis for passive remote sensing of wind speed, and Abbott and Chelton [135] review literature in this field.

Active painting of the sea surface with radar can provide important information on winds. Using microwave scattering techniques, it is possible to estimate near-surface wind velocity under all weather conditions [136]. In particular, the Seasat-A Satellite Scatterometer (SASS) was designed to provide an accuracy of ± 2 m/s and ± 20° direction over a range of 4–26 m/s wind speed. The underlying principle is that very short gravity waves, which affect the strength of Bragg scattering, are in equilibrium with near-surface wind speed. Several wind scatterometers have now been flown in space by NASA, ESA, and NASDA. The first scatterometer was the Seasat Scatterometer (SASS), launched in 1978, but again Seasat had a very short life. The ESA European Remote-Sensing Satellite ERS-1, launched in 1991, carried a scatterometer named the Advanced Microwave Instrument (AMI) scatterometer. This was followed by the ERS-2 AMI scatterometer, launched in 1995. In 1996, NASA launched the NASA Scatterometer (NSCAT), which unfortunately also had a short lifespan. NASA then launched the first scanning scatterometer, "SeaWinds," on QuikSCAT in 1999. A second SeaWinds instrument was flown on the NASDA ADEOS-2 satellite in 1993. The only sensor of this class still operating by mid-2011 was the ASCAT sensor, launched by ESA in 2007.

Future Directions

Historical oceanographic satellite sensors have limited utility in studying or monitoring coastal zones, in part because of their coarse ground resolution and limited spectral resolution and range. Refining ground resolution, expanding the spectral resolution and range, and addressing significant absorbing aerosol contamination issues together create an enormous challenge for accurately distinguish coastal ocean components and characteristics from remote sensing imagery. The limited spectral measurements from the current suite of ocean color sensors are clearly inadequate for coastal zone remote sensing research. Coastal remote sensing presents significant technological challenges. A range of space-based observations, suborbital systems, and models need to be developed over the next 25 years to advance the understanding of coastal habitats. Foremost is the requirement to obtain frequent and synoptic observations of small-scale phenomena in both aquatic and adjacent land environments. The effective discrimination of biogeochemical constituents of the water and seafloor (e.g., colored dissolved organic matter, phytoplankton concentration and composition, suspended sediments, bottom type) and physical properties (e.g., temperature, salinity, wind, circulation, bathymetry, light attenuation) must be achieved over long-term and short-term (daily to weekly) periods, at medium spatial resolution (10–100 m), and within the topographic and bathymetric regime of coastal habitats (watershed to about 20 m depth). Model development must proceed in parallel and at equivalent scales to the new observations. Together these advanced capabilities will lead to new understandings on linkages between lower and higher trophic levels and assessing coastal and estuarine waters for safe ocean resource use and for recreation and cultural purposes. Developing a workforce capable of processing and using these advanced observing technologies and products is also critical. Significant progress can be made by establishing effective links between research and decision-support tools for coastal managers and policymakers.

The next two decades, between 2010 and 2030, will bring a new series of advanced sensors. In the USA, a series of missions is planned. The operational JPSS missions will carry a number of ocean-observing systems, with the precursor NPP satellite scheduled for launch in 2011. The NPP and JPSS will carry the Visible/Infrared Imager/Radiometer Suite (VIIRS). The VIIRS features 22 channels, based primarily on the heritage from three instruments, the NOAA AVHRR, the NASA MODIS, and the Defense Meteorological Satellite Program (DMSP) Operational Linescan System (OLS). VIIRS effectively provides some of the MODIS and SeaWiFS 1-km resolution bands and adds a number of ∼300 m resolution "imaging" channels. Unfortunately, the VIIRS suite of channels, however, does not include the red bands required to estimate solar-stimulated fluorescence that are present on the MODIS and MERIS sensor. As mentioned earlier, these bands are essential to differentiate between phytoplankton blooms and CDOM patches in the coastal ocean.

The Landsat Data Continuity Mission (LDCM), planned for launch in late 2012 or 2013, continues the nearly 40-year legacy of the Landsat satellite series. The LDCM, however, includes a new band, centered at 443 nm, with a lower band edge at 433 nm and an upper band edge at 453 nm, and 30 m resolution. This band will revolutionize coastal water quality applications of the Landsat series. Nevertheless, Landsat-class sensors (NASA's Landsat, France's SPOT, Space Imaging, Inc.'s Ikonos, GEOEye's Quikbird, etc.) provide higher spatial resolution but have limited sensitivity and are unable to detect the subtle changes in reflectance linked to the geophysical properties of interest.

Designing decision-support systems that include synoptic and frequent high-resolution satellite observations to assess coastal and ocean human health is a tractable problem. This will require scientific advances in coastal aquatic environments, innovative techniques, and diverse approaches. An effective strategy must incorporate active and multispectral passive observations and both global and local measurement capabilities. Advanced sensors require expanded capabilities beyond the JPSS and NPP missions, with an aim toward high quality, high spatial, temporal, and spectral observation. Coastal observing capabilities require sensors that can dwell on dark targets or illuminate them with advanced Lidar. Observations need to be consistent and calibrated, accessible, and well-documented.

A series of relevant US missions that provide a solution is being planned following the recommendations of the National Research Council "Decadal Survey" [137]. The HyspIRI, GEO-CAPE, and ACE are a highly complementary set of missions that will provide a wide range of opportunities for multidisciplinary, international collaboration aimed at detecting and monitoring marine pollution. Each of these hyperspectral mission offers an important complementary set of capabilities. Specifically, HyspIRI will have relatively high spatial resolution (60 m), ACE will feature global coverage at medium spatial resolution, and NASA is considering the development of a Pre-ACE mission with similar characteristics, to be launched before 2020, to minimize the potential impacts of a gap in global ocean color observations. GEO-CAPE will afford high temporal coverage from a geostationary orbit. More specifically:

HyspIRI: The NASA HyspIRI (Hyperspectral Infrared Imager) is planned as a polar-orbiter mission with a mid-morning equatorial crossing time and which includes two instruments mounted on a satellite in low Earth orbit. There is an imaging spectrometer measuring from the visible to short-wave infrared (VSWIR) and a multispectral thermal infrared (TIR) imager. The VSWIR and TIR instruments will both have a spatial resolution of 60 m at nadir. The VSWIR will have a temporal revisit of approximately 3 weeks, and the TIR will have a temporal revisit of approximately 1 week. HyspIRI is designed to address issues on both land and in aquatic environments including inland water bodies and shallow coastal environments. The science team helping design this mission seeks to use HyspIRI to examine changes in ecosystem functioning due to many factors including pests, diseases, invasive species, disturbance, climate change, and land management that can alter water quality and food services to humans.

ACE and PACE: The NASA Aerosol-Cloud-Ecology (ACE) Mission and its precursor (the pre-ACE or PACE mission) also are planned polar-orbiter missions with mid-morning equatorial crossing time. These are aerosol-cloud and ocean ecosystem research missions designed to reduce the uncertainty in climate forcing in aerosol-cloud interactions and ocean ecosystem CO_2 uptake. The marine ecosystem goals are to characterize and quantify changes in the ocean biosphere and quantify the amount of dissolved organic matter, carbon, and other biogeochemical species to define the role of the oceans in the carbon cycle (e.g., uptake and storage). The present concepts for the ACE mission includes a wide array of sensors, including Lidars, cloud radars, a multi-angle swath polarimeter for imaging aerosols and clouds, and an ocean color radiometer.

GEO-CAPE: The NASA Geostationary Coastal and Air Pollution Events (GEO-CAPE) mission was recommended to gather science that identifies human versus natural sources of aerosols and ozone precursors, tracks air pollution transport, and studies the dynamics of coastal ecosystems, river plumes, and tidal fronts. The geostationary vantage point would provide a tool with which to examine coastal events in high spatial, (~300 m), high spectral, and high temporal resolution.

The NRC [137] Decadal Survey placed the HyspIRI, GEO-CAPE, and ACE missions in a tier 2 (or phase 2) category, delayed until after a series of other missions deemed prioritary in the report are launched by the US government. Launch for these tier 2 missions is now slated for the second half of the 2010 decade or in the early 2020 decade. It is important to find ways to accelerate this timetable to implement a robust set of tools to monitor and help bridle problems in marine pollution.

It is imperative that the scientific base is strengthened to maximize the application of this technology. As part of this effort, it is important to invest in the training of scientists in the understanding of remote sensing and large multidisciplinary databases. Such scientists will contribute significantly to advances and conscious policymaking regarding marine pollution.

Bibliography

1. Muller-Karger FE (1992) Remote sensing of marine pollution: a challenge for the 1990s. Mar Pollut Bull 25(1–4):54–60
2. Boyce DG, Lewis MR, Worm B (2010) Global phytoplankton decline over the past century. Nature 466:591–596. doi:10.1038/nature09268
3. Polovina JJ, Howell EA, Abecassis M (2008) Ocean's least productive waters are expanding. Geophys Res Lett 35: L03618. doi:10.1029/2007GL031745
4. Behrenfeld MJ, O'Malley RT, Siegel DA, McClain CR, Sarmiento JL, Feldman GC, Milligan AJ, Falkowski PG, Letelier RM, Boss ES (2006) Climate-driven trends in contemporary ocean productivity. Nature 444:752–755 (7 Dec 2006). doi:10.1038/nature05317
5. Gledhill DK, Wanninkhof R, Eakin CM (2009) Observing ocean acidification from space. Oceanography 22(4):48–59
6. Orr JC et al (2005) Anthropogenic ocean acidification over the twenty-first century and its impact on calcifying organisms. Nature 437(7059):681–686
7. Chavez FP, Ryan J, Lluch-Cota SE, Miguel ÑC (2003) From anchovies to sardines and back: multidecadal change in the Pacific ocean. Science 299:217
8. Polovina JJ (2005) Climate variation, regime shifts, and implications for sustainable fisheries. Bull Mar Sci 76(2): 233–244
9. Epstein PR (1999) Climate and health. Perspective. Science 285(5426):347–348. doi:10.1126/science.285.5426.347
10. Zirino A, Fiedler PC, Keir RS (1988) Surface pH, satellite imagery, and vertical models in the tropical ocean. Sci Total Environ 75:285–300
11. Dube C, Lamarche A, Alfoldi T (1989) Resultats preliminaires d'une methode d'evaluation de la dispersion des rejets des eaux usees dans le Fleuve Saint-Laurent par teledetection. In: IGARSS proceedings: quantitative remote sensing: an economic tool for the nineties, 12th Canadian symposium on remote sensing, IEEE, Vancouver, Canada, vol 5: pp2820–2824
12. Gordon HR, McCluney WR (1975) Estimation of the depth of sunlight penetration in the sea for remote sensing. Appl Opt 14:413–416
13. Boyd JD, Myrick RK, Linzell RS (1991) Isis: a portable system for near real time oceanographic analysis. In: Abstract, EOS Supplement, AGU 1992 Ocean sciences meeting, New Orleans 72(51):28
14. Frye DE, Fougere A, Kery S (1991) Prototype expendable surface mooring with inductive telemetry. In: Abstract, EOS Supplement, AGU 1992 Ocean sciences meeting, New Orleans 72(51):29
15. OSB (Ocean Studies Board) (2003) Enabling ocean research in the 21st century: implementation of a network of ocean observatories. National Research Council. The National Academies Press, Washington, DC, 221p
16. Walt D, Urban E (1991) Chemical measurement technologies for ocean science. In: Abstract, EOS Supplement, AGU 1992 Ocean sciences meeting, New Orleans 72(51):28
17. Whitledge TE, Liljestrand HM (1991) In situ nitrate analyzer: design, development and field results. In: Abstract, EOS Supplement, AGU 1992 Ocean sciences meeting, New Orleans 72(51):27
18. Johnson KS, Needoba JA, Riser SC, Showers WJ (2007) Chemical sensor networks for the aquatic environment. Am Chem Soc. doi:10.1021/cr050354e
19. Honda MC, Watanabe S (2007) Utility of an automatic water sampler to observe seasonal variability in nutrients and DIC in the northwestern North Pacific. J Oceanogr 63:349–362
20. McPhaden MJ, Milburn HB (1991) Moored precipitation measurements. In: Abstract, EOS Supplement, AGU 1992 Ocean sciences meeting, New Orleans 72(51):27
21. Serra YL, McPhaden MJ (2004) In situ observations of diurnal variability in rainfall over the tropical Pacific and Atlantic oceans. J Climate 17:3496–3509
22. Case JF, Widder EA (1991) HIDEX-type bioluminescence detectors: modifications for moored and towed use. In: Abstract, EOS Supplement, AGU 1992 Ocean sciences meeting, New Orleans 72(51):30
23. Lapota D, Geiger ML, Lavoie DM, Bernstein SH, Case JF (1991) Measurements of planktonic bioluminescence in Vestfjord, Norway using HIDEX, a new rapid profiling bathyphotometer. In: Abstract, EOS Supplement, AGU 1992 Ocean sciences meeting, New Orleans 72(51):30
24. Widder EA (2010) Bioluminescence in the ocean: origins of biological, chemical, and ecological diversity. Science 328(5979):704–708

25. Jaffe JS (1991) Three dimensional sonar sensing of underwater animals. In: Abstract, EOS Supplement, AGU 1992 Ocean sciences meeting, New Orleans 72(51):28
26. Macaulay MC (1991) Applications of hydroacoustic technology to the study of zooplankton and micronekton in open ocean and shallow water environments. In: Abstract, EOS Supplement, AGU 1992 Ocean sciences meeting, New Orleans 72(51):28
27. Remsen AW (2008) Evolution and field application of a plankton imaging system. Ph.D. dissertation, University of South Florida, 145p
28. Smith RC, Waters KJ, Baker KS (1991) Optical variability and pigment biomass in the Sargasso Sea as determined using deep-sea optical mooring data. J Geophys Res 96(C5): 8665–8686
29. Dickey T, Marra J, Granata T, Langdon C, Hamilton M, Wiggert J, Siegel D, Bratkovich A (1991) Concurrent high resolution bio-optical and physical time series observations in the Sargasso Sea during the spring of 1987. J Geophys Res 96(C5):8643–8663
30. Boss E, Behrenfeld M (2010) In situ evaluation of the initiation of the North Atlantic phytoplankton bloom. Geophys Res Lett 37(L18603):5
31. Dickey TD, Itsweire EC, Moline MA, Perry MJ (2008) Introduction to the limnology and oceanography special issue on autonomous and lagrangian platforms and sensors (ALPS). Limnol Oceanogr 53(5, part 2):2057–2061
32. Esaias WE (1980) Remote sensing of oceanic phytoplankton: present capabilities and future goals. In: Falkowski PG (ed) Primary productivity in the sea. Plenum, New York, pp 321–337
33. Freilich MH (2010) NASA Earth science activities related to the deepwater horizon oil spill. American Geophysical Union, Fall Meeting 2010, Abstract #U14A-02
34. Zielinski O, Busch JA, Cembella AD, Daly KL, Engelbrektsson J, Hannides AK, Schmidt H (2009) Detecting marine hazardous substances and organisms: sensors for pollutants, toxins, and pathogens. Ocean Sci Discuss 6:953–1005. www.ocean-sci-discuss.net/6/953/2009/
35. Chomko RM, Gordon HR (2001) Atmospheric correction of ocean color imagery: test of the spectral optimization algorithm with the sea-viewing wide field-of-view sensor. Appl Opt 40:2973–2984
36. Chomko RM, Gordon HR, Maritorena S, Siegel DA (2003) Simultaneous retrieval of oceanic and atmospheric parameters for ocean color imagery by spectral optimization: a validation. Remote Sens Environ 84:208–220
37. Hovis WA, Clark DK, Anderson F, Austin RW, Wilson WH, Baker ET, Ball D, Gordon HR, Mueller JL, El-Sayed SZ, Sturm B, Wrigley RC, Yentsch CS (1980) Nimbus-7 coastal zone color scanner: system description and initial imagery. Science 210:60–63
38. Feldman G, Kuring N, Ng C, Esaias WE, McClain C, Elrod J, Maynard N, Endres D, Evans R, Brown J, Walsh S, Carle M, Podesta G (1989) Ocean color, availability of the global data set. The oceanography report. EOS 70(23):634
39. Barale V, Rizzoli PM, Hendershott MC (1984) Remotely sensing the surface dynamics of the Adriatic Sea. Deep Sea Res 31(12):1433–1459
40. Barale V, McClain CR, Rizzoli PM (1986) Space and time variability of the surface color field in the northern Adriatic Sea. J Geophys Res 91(C11):12957–12974
41. Bergamasco A, Barale V (1988) Comparison between coastal runoffs patterns from CZCS imagery and from a general circulation model. In: Marani A (ed) Advances in environmental modelling. Elsevier, Amsterdam, pp 395–404
42. Holligan PM, Aarup T, Groom SB (1989) The North Sea satellite colour atlas. Cont Shelf Res 9(8):667–765
43. Muller-Karger FE, McClain CR, Richardson PL (1988) The dispersal of the Amazon's water. Nature 333:56–59
44. Muller-Karger FE, McClain CR, Fisher TR, Esaias WE, Varela R (1989) Pigment distribution in the Caribbean Sea: observations from space. Prog Oceanogr 23:23–69
45. Müller-Karger FE, Walsh JJ, Evans RH, Meyers MB (1991) On the seasonal phytoplankton concentration and sea surface temperature cycles of the Gulf of Mexico as determined by satellites. J Geophys Res 96(C7):12645–12665
46. Del Castillo C, Gilbes F, Coble P, Muller-Karger FE (2000) On the dispersal of riverine colored dissolved organic matter over the West Florida Shelf. Limnol Oceanogr 45(6):1425–1432
47. IOCCG (2000) Remote sensing of ocean colour in coastal, and other optically-complex waters. In: Sathyendranath S (ed) International Ocean-Colour Coordinating Group, No.3: general introduction. International Ocean-Colour Coordinating Group (IOCCG) Report, pp 5–22
48. Abbott M, Letelier R (1999) Chlorophyll fluorescence (MODIS Product No. 20). Algorithm theoretical basis document. http://modis.gsfc.nasa.gov/data/atbd/atbd_mod22.pdf. Accessed 11 October 2011
49. SWFDOG (2002) Satellite images track 'black water' event off Florida coast. EOS Trans AGU 83(281):285
50. Hu C, Hackett KE, Callahan MK, Andréfouët S, Wheaton JL, Porter JW, Muller-Karger FE (2003) The 2002 ocean color anomaly in the Florida Bight: a cause of local coral reef decline? Geophys Res Lett 30(3):1151. doi:10.1029/2002GL016479
51. Hu C, Muller-Karger FE, Taylor C, Myhre D, Murch B, Odriozola AL, Godoy G (2003) MODIS detects oil spills in Lake Maracaibo, Venezuela. Eos Trans Am Geophys Union 84(33):313–319
52. Gordon HR, Clark DK, Brown JW, Brown OB, Evans RH, Broenkow WW (1983) Phytoplankton concentrations in the Middle Atlantic Bight: comparison ship determinations and CZCS estimates. Appl Opt 22:20–35
53. O'Reilly JE et al (2000) Ocean chlorophyll-a algorithms for SeaWiFS, OC2 and OC4: Version 4. In: Hooker SB, Firestone ER (eds) SeaWiFS postlaunch calibration and validation analyses, Part 3. NASA Tech. Memo. 2000–206892, vol 11. NASA Goddard Space Flight Center, Greenbelt, pp 9–23
54. Carder KL, Chen FR, Lee ZP, Hawes S, Kamykowski D (1999) Semi-analytic MODIS algorithms for chlorophyll a and absorption with bio-optical domains based on nitrate-depletion temperatures. J Geophys Res 104(C3):5403–5421

55. Hu C, Lee ZP, Muller-Karger FE, Carder KL (2002) Application of an optimization algorithm to satellite ocean color imagery: a case study in Southwest Florida coastal waters. In: Frouin RJ, Yuan Y, Kawamura H (eds) Ocean remote sensing and applications: SPIE proceedings, Washington, USA, vol 4892, pp 70–79
56. Maritorena S, Siegel DA, Peterson AR (2002) Optimization of a semianalytical ocean color model for global-scale applications. Appl Opt 41:2705–2714
57. Lee Z, Carder KL, Arnone RA (2002) Deriving inherent optical properties from water color: a multiband quasi-analytical algorithm for optically deep waters. Appl Opt 41:5755–5772
58. Morel A, Gentili B (2009) A simple band ratio technique to quantify the colored dissolved and detrital organic material from ocean color remotely sensed data. Remote Sens Environ 113:998–1011
59. Siegel DA, Maritorena S, Nelson NB, Behrenfeld MJ (2005) Independence and interdependencies of global ocean color properties: re-assessing the bio-optical assumption. J Geophys Res 110:C07011. doi:10.1029/2004JC002527
60. Chen Z, Muller-Karger FE, Hu C (2007) Remote sensing of water clarity in Tampa Bay. Remote Sens Environ. doi:10.1016/j.rse.2007.01.002
61. Stumpf RP (2001) Applications of satellite ocean color sensors for monitoring and predicting harmful algal blooms. Hum Ecol Risk Assess 7:1363–1368
62. Hu C, Muller-Karger FE, Taylor C, Carder KL, Kelble C, Johns E, Heil CA (2005) Red tide detection and tracing using MODIS fluorescence data: a regional example in SW Florida coastal waters. Remote Sens Environ 97:311–321
63. Hu C, Cannizzaro J, Carder KL, Muller-Karger FE, Hardy R (2010) Remote detection of Trichodesmium blooms in optically complex coastal waters: examples with MODIS full-spectral data. Remote Sens Environ 114(9):2048–2058. doi:10.1016/j.rse.2010.04.011
64. Hu C, Li D, Chen C, Ge J, Muller-Karger FE, Liu J, Yu F, He M-X (2010) On the recurrent Ulva prolifera blooms in the Yellow Sea and East China Sea. J Geophys Res 115(C5). doi:10.1029/2009JC005561
65. Zhang H (2002) Detecting red tides on the West Florida shelf by classification of SeaWiFS satellite imagery. Master's thesis, Department of Computer Science and Engineering, University of South Florida
66. Cannizzaro et al (2004) Bio-optical signatures of red tides on the west Florida shelf. Cont Shelf Res 28(1):137–158
67. Miller RL, McKee BA (2004) Using MODIS Terra 250 m imagery to map concentrations of total suspended matter in coastal waters. Remote Sens Environ 93(2004):259–266
68. Chen Z, Hu C, Conmy RN, Swarzenski P, Muller-Karger F (2007) Colored dissolved organic matter in Tampa Bay, Florida. Mar Chem 104:98–109
69. Chen Z, Hu C, Muller-Karger FE (2007) Monitoring turbidity in Tampa Bay using MODIS 250 M imagery. Remote Sens Environ. doi:10.1016/j.rse.2006.12.019
70. Moreno MJ, Al-Hamdan M, Rickman D, Muller-Karger FE (2010) Using the surface reflectance MODIS terra product to estimate turbidity in Tampa Bay, Florida. Remote Sens 2(12):2713–2728. doi:10.3390/rs2122713
71. Doxaran D, Froidefond J-M, Castaing P, Babin M (2009) Dynamics of the turbidity maximum zone in a macrotidal estuary (the Gironde, France): observations from field and MODIS satellite data. Estuar Coast Shelf Sci 81:321–332
72. Lobitz B, Beck L, Huq A, Wood B, Fuchs G, Faruque ASG, Colwell R (2000) Climate and infectious disease: use of remote sensing for detection of *Vibrio cholerae* by indirect measurement. Proc Natl Acad Sci USA 97(4):1438–1443
73. Colwell RR (2005) Global climate and infectious disease: the cholera paradigm. Science 274:2025–2031
74. Hu W, Clements A, Williams G, Tong S (2010) Dengue fever and El Niño/Southern Oscillation in Queensland, Australia: a time series predictive model. Occup Environ Med 67:307–311. doi:10.1136/oem.2008.044966
75. Gagliardini DA, Karszenbaum H, Legeckis R, Klemas V (1984) Application of LANDSAT MSS, NOAA/TIROS AVHRR, and Nimbus CZCS to study the La Plata River and its interaction with the ocean. Remote Sens Environ 15:21–36
76. Stumpf RP, Tyler MA (1988) Satellite detection of bloom and pigment distributions in estuaries. Remote Sens Environ 24:385–404
77. Froidefond JM, Castaing P, Jouanneau JM, Prud'Homme R, Dinet A (1993) Method for the quantification of suspended sediments from AVHRR NOAA-11 satellite data. Int J Remote Sens 14(5):885–894
78. Maul GA (1985) Introduction to satellite oceanography. Martinus Nijhoff, Dordrecht/Boston, 606p
79. Maul GA, Gordon HR (1975) On the use of the Earth resources technology satellite (LANDSAT-1) in optical oceanography. Remote Sens Environ 4:95–128
80. Amos CL, Alfoldi TT (1979) The determination of suspended sediment concentration in a macrotidal system using Landsat data. J Sediment Petrol 49:159–174
81. Khorram S (1981) Water quality mapping from Landsat digital data. Int J Rem Sens 2(2):145–153
82. Dwivedi RM, Narain A (1987) Remote sensing of phytoplankton: An attempt from the landsat thematic mapper. Int J Remote Sens 8(10):1563–1569
83. Braga CZF, Setzer AW, Drude de Lacerda L (1993) Water quality assessment with simultaneous Landsat-5 TM data at Guanabara Bay, Rio de Janeiro, Brazil. Remote Sens Environ 45:95–106
84. Tassan S (1987) Evaluation of the potential of the tematic mapper for marine application. Int J Remote Sens 8(10):1455–1478
85. Tassan S (1993) An improved in-water algorithm for the determination of chlorophyll and suspended sediment concentration from thematic mapper data in coastal waters. Int J Remote Sens 14(6):1221–1229
86. Khorram S, Cheshire H, Geraci AL, La Rosa G (1989) IGARSS proceedings: quantitative remote sensing: an economic tool for the nineties, 12th Canadian symposium on remote sensing, IEEE, Vancouver, Canada, vol 1, pp 335–338

87. Catts GP, Khorram S, Cloern JE, Knight AW, DeGloria SD (1985) Remote sensing of tidal chlorophyll a variations in estuaries. Int J Remote Sens 6(11):1685–1706
88. Munday JC, Fedosh MS (1981) Chesapeake Bay plume dynamics from Landsat. In: Campbell JW, Thomas JP (eds) Superflux: Chesapeake Bay plume study, NASA Conference Publ 2188. NASA, Greenbelt
89. Munday JC, Alfoldi TT (1979) Landsat test of diffuse reflectance models for aquatic suspended solids measurement. Remote Sens Environ 8:169–183
90. Hellweger FL, Miller W, Oshodi KS (2007) Mapping turbidity in the Charles River, Boston using a high-resolution satellite. Environ Monit Assess 132(1–3):311–320, Epub 14 Dec 2006
91. Sorensen K, Nilsen J, Saebo HV, Holbaek-Hanssen E (1989a) Use of thematic mapper data for mapping of water quality. In: IGARSS proceedings: quantitative remote sensing: an economic tool for the nineties, 12th Canadian symposium on remote sensing, IEEE, Vancouver, Canada, vol 2, p 696
92. Sorensen K, Lindell T, Nisell J (1989b) The information content of AVHRR, MSS, TM, and SPOT data in the Skagerrak Sea. In: IGARSS proceedings: quantitative remote sensing: an economic tool for the nineties, 12th Canadian symposium on remote sensing, IEEE, Vancouver, Canada, vol 4, pp 2439–2442
93. Lee Z, Carder KL, Mobley CD, Steward RG, Patch JS (1999) Hyperspectral remote sensing for shallow waters. II. Deriving bottom depths and water properties by optimization. Appl Opt 38:3831–3843
94. Brando VE, Dekker AG (2003) Satellite hyperspectral remote sensing for estimating estuarine and coastal water quality. IEEE Trans Geosci Remote Sens 41(6):1378–1387
95. Corson MR, Korwan DR, Lucke RL, Snyder WA, Davis CO (2008) The hyperspectral imager for the coastal ocean (HICO) on the international space station. In: IEEE proceedings of the international geoscience and remote sensing symposium, Boston, USA, 978-1-4244-2808-3/08
96. Vane G, Chrisp M, Enmark H, Macenka S, Solomon J (1984a) Airborne visible/infrared imaging spectrometer: an advanced tool for earth remote sensing. In: IGARSS '84, Strasbourg, France, SP215, p 751
97. Vane G, Goetz AFH, Wellman JB (1984) Airborne imaging spectrometer: a new tool for remote sensing. IEEE Trans Geosci Remote Sens GE-22:546
98. Goetz AF (1987) High-resolution imaging spectrometer: science opportunities for the 1990s. NASA Earth observing system report: instrument panel report, vol IIc, 74p
99. Karaska MA, Huguenin RL, Beacham JL, Wang Mo-Hwa, Jensen JR, Kaufmann RS (2004) AVIRIS measurements of chlorophyll, suspended minerals, dissolved organic carbon, and turbidity in the Neuse River, North Carolina. Photogramm Eng Remote Sensing 70(1):125–133
100. Bagheri S, Yu T (2008) Hyperspectral sensing for assessing nearshore water quality conditions of Hudson/Raritan estuary. J Environ Inf 11(2):123–130. doi:10.3808/jei.200800116
101. Gower JFR, Buxton RAH, Borstad GA (1989) The FLI airborne imaging spectrometer: experience with land and water targets. In: IGARSS proceedings: quantitative remote sensing: an economic tool for the nineties, 12th Canadian symposium on remote sensing, IEEE, Vancouver, Canada, vol 2, pp 1024–102
102. Dekker AG, Malthus TJ, Seyhan E (2002) Quantitative modelling of inland water quality for high resolution MSS-systems. Geosci Remote Sens IEEE Trans 29(1):89–95
103. Babey SK, Anger CD (1989) A compact airborne spectrographic imager (CASI). In: IGARSS proceedings: quantitative remote sensing: an economic tool for the nineties, 12th Canadian symposium on remote sensing, IEEE, Vancouver, Canada, vol 2, p 1028–1031
104. Ammenberg P, Flink P, Lindell T, Pierson D, Strombeck N (2002) Bio-optical modelling combined with remote sensing to assess water quality. Int J Remote Sens 23(8): 1621–1638
105. Hengstermann T, Reuter R (1990) Lidar fluorosensing of mineral oil spills on the sea surface. Appl Opt 29(22):3218–3227
106. Hoge FE, Swift RN (1980) Application of the NASA airborne oceanographic lidar to the mapping of chlorophyll and other organic pigments. In: Campbell JW, Thomas JP (eds) Superflux: Chesapeake Bay Plume Study, Conference Publ. 2188. NASA, Greenbelt, pp 349–374
107. Hoge FE, Swift RN (1983) Airborne detection of oceanic turbidity cell structure using depth-resolved laser-induced water Raman backscatter. Appl Opt 23:3778–3786
108. Brekke C, Solberg AHS (2005) Oil spill detection by satellite remote sensing. Remote Sens Environ 95:1–13
109. Garcia-Pineda O, MacDonald I, Zimmer B, Shedd B, Roberts H (2010) Remote-sensing evaluation of geophysical anomaly sites in the outer continental slope, northern Gulf of Mexico. Deep Sea Res Part 2 Top Stud Oceanogr 57:1859–1869
110. Korenowski GM, Frysinger GS, Asher WE, Barger WR, Klusty MA (1989) Laser based optical measurement of organic surfactant concentration variations a te air/sea interface. In: IGARSS proceedings: quantitative remote sensing: an economic tool for the nineties, 12th Canadian symposium on remote sensing, IEEE, Vancouver, Canada, vol 3, pp 1506–1509
111. Hoge FE, Swift RN (1980) Oil film thickness measurement using airborne laser-induced water Raman backscatter. Appl Opt 19(19):3269–3281
112. Hoge FE, Swift RN (1983) Experimental feasibility of the airborne measurement of absolute oil fluorescence spectral conversion efficiency. Appl Opt 22(1):37–47
113. Hoge F, Swift RN (1982) Delineation of estuarine fronts in the German Bight using airborne laser-induced water Raman backscatter and fluorescence of water column constituents. Int J Remote Sens 3:475–495
114. Vodacek A (1989) Synchronous fluorescence spectroscopy of dissolved organic matter to optimize lidar detection parameters. In: IGARSS proceedings: quantitative remote sensing: an economic tool for the nineties, 12th Canadian symposium on remote sensing, IEEE, Vancouver, Canada, vol 2, pp 1046–1049

115. Duncan ME, Ackleson SG (1991) A summary of hand-held photography of the Persian Gulf area taken during space shuttle missions: 1981–1991. In: Abstract, EOS Supplement, AGU 1992 Ocean sciences meeting, New Orleans 72(51):40
116. Raney RK (1983) The Canadian SAR experience, Chapter 13. In: Allan TD (ed) Satellite microwave sensing. Wiley, Tornoto, pp 223–234
117. Wadsworth A, Robertson C, De Staerke D (1983) Use of SEASAT-SAR data in oceanography at the IFP, Chapter 14. In: Allan TD (ed) Satellite microwave sensing. Ellis Horwood Ltd , Chichester pp 235–245
118. Tucker C, Holben B, Elgin J, McMurtrey J (1989) Relationship of spectral data to grain yield variation. Photogramm Eng Remote Sensing 46(5):657–666
119. Tucker CJ, Sellers PJ (1986) Satellite remote sensing of primary production. Int J Remote Sens 7:1395–1416
120. Hu C, Muller-Karger F, Murch B, Myhre D, Taylor J, Luerssen R, Moses C, Zhang C, Gramer L, Hendee J (2009) Building an automated integrated observing system to detect sea surface temperature anomaly events in the Florida Keys. IEEE Trans Geosci Remote Sens 47(6):1607–1620
121. Barnes BB, Chuanmin Hu, Muller-Karger F (2011) An improved high-resolution SST climatology to assess cold water events off Florida. Geosci Remote Sens Lett 8:769, (Accepted Jan 2011)
122. Soto I (2006) Environmental variability in the Florida keys: impacts on coral reef health. Master's thesis, University of South Florida, College of Marine Science
123. Eakin CM, Nim CJ, Brainard RE, Aubrecht C, Elvidge C, Gledhill DK, Muller-Karger F, Mumby PJ, Skirving WJ, Strong AE, Wang M, Weeks S, Wentz F, Ziskin D (2010) Monitoring coral reefs from space. Oceanogr Soc Mag, special volume: The future of oceanography from space, Dec 2010, pp 119–133
124. Thomas A, Byrne D, Weatherbee R (2002) Coastal sea surface temperature variability from Landsat infrared data. Remote Sens Environ 81:262–272
125. Fisher JI, Mustard JF (2004) High spatial resolution sea surface climatology from Landsat thermal infrared data. Remote Sens Environ 90:293–307
126. Lagerloef GSE, Swift CT, Levine DM (1991) Remote sensing of sea surface salinity: airborne and satellite concepts. In: Abstract, EOS Supplement, AGU 1992 Ocean sciences meeting, New Orleans 72(51):29
127. Wagner CA, Cheney RE (1992) Global sea level change from satellite altimetry. J Geophys Res 97(C10):15607–15615. doi:10.1029/92JC01641
128. Merrifield MA, Merrifield ST, Mitchum GT (2010) Evidence for anomalous recent acceleration of global sea level rise. J Climate 22:5772–5781
129. Bindoff NL, Willebrand J, Artale V, Cazenave A, Gregory J, Gulev S, Hanawa K, Le Quéré C, Levitus S, Nojiri Y, Shum CK, Talley LD, Unnikrishnan A (2007) Observations: oceanic climate change and sea level. In: Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds) Climate change 2007: the physical science basis. Contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge/New York
130. Emery WJ, Strub T, Leben R, Foreman M, McWilliams JC, Han G, Ladd C, Ueno H (2010) Satellite altimetry applications off the coasts of North America. In: Vignudelli S, Kostianoy A, Cipollini P, Benveniste J (eds) Coastal altimetry. Springer Verlag, Germany, pp 417–451. doi:10.1007/978-3-642-12796-0_16
131. Hwang PA, Teague WJ, Jacobs GA, Wang DW (1998) A statistical comparison of wind speed, wave height and wave period derived from satellite altimeters and ocean buoys in the Gulf of Mexico Region. J Geophys Res 103:10451–10468
132. Hwang PA, Walsh EJ, Krabill WB, Swift RN, Manizade SS, Scott JF, Earle MD (1998) Airborne remote sensing applications to coastal wave research. J Geophys Res 103:18791–18800
133. Bidlot J-R, Holmes DJ, Wittmann PA, Lalbeharry R, Chen HS (2002) Intercomparison of the performance of operational ocean wave forecasting systems with buoy data. Weather Forecasting 17:287–310. doi:10.1175/1520-0434(2002)017<0287:IOTPOO>2.0.CO;2
134. Swift CT (1990) Passive microwave remote sensing of ocean surface wind speed. In: Geernaert GL, Plant WJ (eds) Surface waves and fluxes, vol 2. Kluwer Academic, Dordrecht, pp 265–292
135. Abbott MR, Chelton DB (1991) Advances in passive remote sensing of the ocean. Contributions in Oceanography. U.S. National Report to International Union of Geodesy and Gophysics 1987–1990. American Geophysical Union, p 571–589
136. Guymer TH (1983) Validation and applications of SASS over JASIN, Chapter 5. In: Allan TD (ed) Satellite microwave sensing. Wiley, Toronto, pp 87–104
137. NRC (2007) Earth science and applications from space: national imperatives for the next decade and beyond. Committee on Earth science and applications from space: a community assessment and strategy for the future. National Research Council, p 456. ISBN: 978-0-309-10387-9
138. Chadwick DB, Lieberman SH, Reimers CE (1991) In-situ release rate measurements of contaminants from marine sediments. In: Abstract, EOS Supplement, AGU 1992 Ocean sciences meeting, New Orleans 72(51):29
139. Esaias W (1986) MODIS – moderate resolution imaging spectrometer. NASA Earth observing system instrument panel report, vol IIb, 59p
140. Huang WG, Lou XL (2003) AVHRR detection of red tides with neural networks. Int J Remote Sens 24:1991–1996
141. Muller-Karger FE, Hu C, Andréfouët S, Varela R (2005) The color of the coastal ocean and applications in the solution of research and management problems. In: Miller RL, Del Castillo CE, McKee BA (eds) Remote sensing of coastal aquatic environments: technologies, techniques and application. Springer, Dordrecht, pp 101–127
142. Nerem RS, Leuliette E, Cazenave A (2006) Present-day sea-level change: a review. Comptes Rendus Geoscience 338:1077–1083

143. Stumpf RP (1988) Sediment transport in chesapeake bay during floods: analysis using satellite and surface observations. J Coast Res 4(1):1–15
144. McClain, Charles R, Cleave ML, Feldman GC, Gregg WW, Hooker SB, Kuring N (1998) Science quality seaWiFS data for global biosphere research. Sea Technology. September 1998, pages 10–16
145. Rodríguez-Guzmán V, Gilbes F (2009) Estimating total suspended sediments in tropical open bay conditions using MODIS. In: Proceedings of the 8th WSEAS International Conference on Instrumentation, Measurement, Circuits and Systems, Hangzhou, China, May 20–22, 2009, pp 83–86
146. Gallegos SC, Gray TI, Crawford MM (1989) A study into the responses of the NOAA-n AVHRR reflective channels over water targets. In: Proceedings of the 1989 IEEE IGARSS Meeting, Vancouver, BC
147. Stumpf RP, Pennock JR (1991) Remote estimation of the diffuse attenuaton coefficient in a moderately turbid estuary. Rem. Sens. Environ. 38:183–191
148. Fingas MF, Brown CE (1997) Airborne oil spill remote sensors - do they have a future. In: Proceedings of the Third International Airborne Remote Sensing Conference and Exhibition, Environmental Research Institute of Michigan (ERIM), Ann Arbor, MI, pp I 715–722
149. Fingas MF, Brown CE (2000) Review of oil spill remote sensing. In: proceedings of the Sixth International Conference on Remote Sensing for Marine and Coastal Environments, Veridian ERIM International, Ann Arbor, MI, pp I211–218
150. Stumpf RP, Pennock JR (1989) Calibration of a general optical equation for remote sensing of suspended sediments in a moderately turbid estuary. J Geophys Res 94 (C10):14363–14371

Remote Sensing of Natural Disasters

Steve Chien[1], Veerachai Tanpipat[2]
[1]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA
[2]Forest Fire Control Division, National Park, Wildlife and Plant Conservation Department of Thailand, Chatuchak, Bangkok, Thailand

Article Outline

Glossary
Definition: Remote Sensing of Natural Disasters
Introduction
Remote Sensing for the Study of Volcanic Activity
Flooding Applications of Remote Sensing
Fire Applications of Remote Sensing
Other Applications of Remote Sensing to Natural Disasters
Future Directions
Conclusions
Acknowledgments
Bibliography

Glossary

MODIS The MODerate resolution Imaging Spectrometer is a general purpose instrument flying on the Terra and Aqua spacecraft that can sense both visible and thermal infrared information on the earth's surface and atmosphere.

GOES The Geostationary Operational Environmental Satellite system is a set of satellites that continuously cover fixed regions of Earth. For example, GOES-East provides continuous coverage of North and South America.

AVHRR Advanced Very High Resolution Radiometer is a sensor carried on the National Oceanic and Atmospheric Administration (NOAA) family of polar orbiting platforms (POES). AVHRR has five wide spectral bands that sense principally the near-infrared and thermal infrared spectrum.

ASTER Advanced Spaceborne Thermal Emission and Reflection Radiometer is an instrument on the Terra satellite that senses visible, near-infrared, short infrared, and long-wave/thermal infrared spectrum.

Earth Observing One (EO-1) An Earth orbiting, pointable spacecraft that has been used to demonstrate a wide range of automation and autonomic technologies including onboard mission replanning and sensorwebs.

Hyperion The Hyperspectral instrument on EO-1 used onboard to detect flooding, volcanic activity, and cryosphere change. Hyperion is able to measure in the Very Near to Short Wave infrared spectrum.

Synthetic aperture radar (SAR) A radar remote sensing technique in which motion of the radar is used to synthesize a large array through radar processing. SAR can be used to distinguish between various surface types such as water, land, vegetation cover type and density, and others.

Interferometric synthetic aperture radar (InSAR) A remote sensing technique which enables detailed topography, change detection, and motion tracking with great precision. InSAR has applications to deriving digital elevation maps (DEMs) as well as tracking land motion (e.g., after earthquakes), ice sheet motion, and change detection after major disturbances such as landslides, fires, flooding, and other natural disasters.

Lahar A mud or debris flow–caused volcanic activity combined with water, snow, or ice.

Definition: Remote Sensing of Natural Disasters

Remote sensing involves the use of instruments to study phenomena from a distance. Natural disasters derive from natural hazards such as volcanoes, flooding, fires, and weather. Practically speaking, "remote sensing of natural disasters" principally refers to the use of airborne or spaceborne sensors to study natural disasters for detecting, modeling, predicting, analyzing, and mitigating effects on human populations and activities.

Remote sensing can be further characterized by the sensitivity of the instrument. Passive sensors rely either on reflected sunlight to illuminate the target or by radiation emitted from the target. Typical passive sensors might include the visible spectrum (~1 μm wavelength) or very near and short wave infrared (up to 2.5 μm wavelength). These spectra tend to be useful for distinguishing various vegetation and surface types. Applications for monitoring of natural hazards include detection of urban areas, plants, surface water, ice, snow, lava, lahar, and roads. Typical passive sensors for emissive targets would include sensitivity in the Thermal Infrared (TIR) spectrum (8–12 μm wavelength). Applications of TIR to natural hazards include measurement of volcanic activity, and active fire mapping.

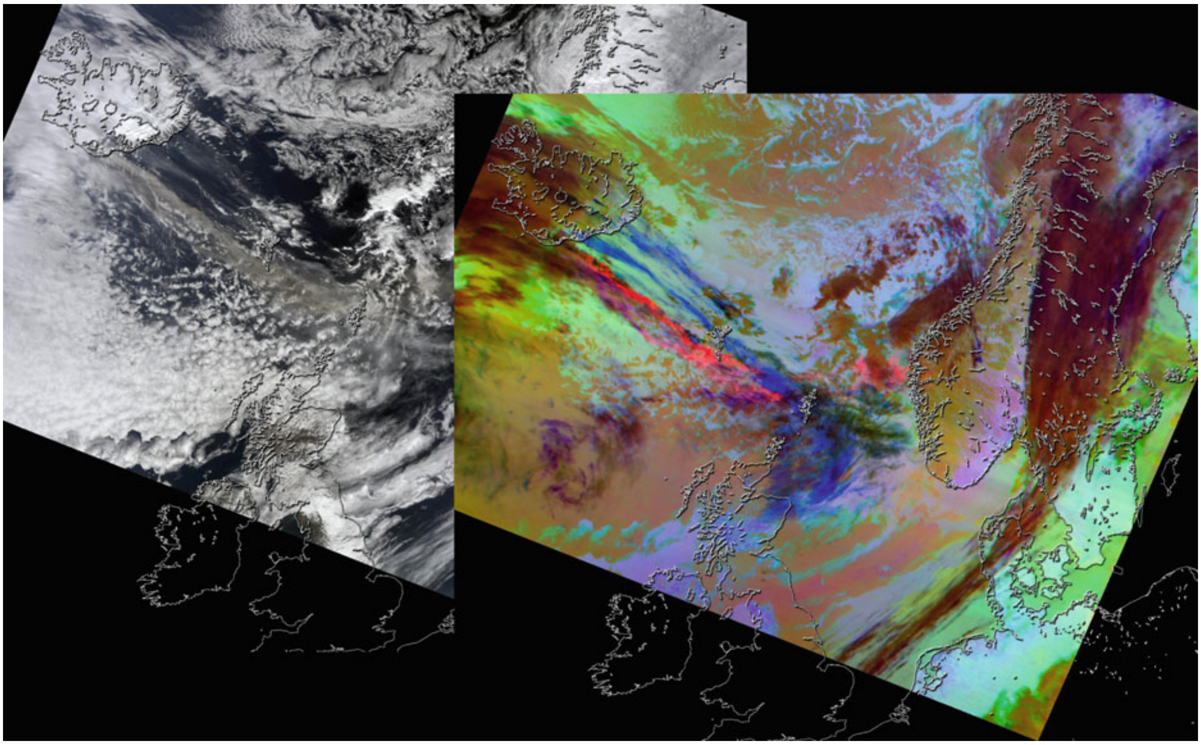
Active sensors such as radar produce a signal to illuminate the surface of interest and rely on the surface impact on the reflected energy to distinguish different surface characteristics. For example, Synthetic Aperture Radar (SAR) emits a signal with specific polarizations (e.g., HH, VV) and measures the returned energy at various polarizations to distinguish surface characteristics. These radars are useful in that they are less affected by weather such as cloud cover and thus have unique capabilities in flooding and fire applications where clouds and smoke may obscure areas of interest.

In this entry we focus on spaceborne applications. However, all of the sensing modalities described (e.g., visible, near-infrared, very near-infrared, thermal infrared, radar, SAR, etc.) are also often provided via airborne platforms. Thus, all of the same applications of space-based remote sensing are relevant to airborne remote sensing.

Introduction

Remote sensing can provide valuable data to aid in the study and mitigation of natural disasters. Satellite instruments can provide imagery of large areas of Earth's surface and overflights can often easily provide data for hard to access locations (either due to terrain, weather, or causes of the disaster itself). Airborne instruments can also provide much of the same utility from remote sensing with the potential advantages of greater loiter or dwell time and rapid response. While in the remainder of this entry we focus on spaceborne remote sensing, generally speaking, any of the applications of remote sensing described can be deployed as airborne instruments as well.

In this entry, we provide an overview of the wide range of methods used to apply remote sensing to the study and mitigation of natural disasters. We first describe the use of remote sensing to study volcanic activity – including measurement of thermal emissions, mapping of lava flows, measurement of plumes and ash, and gas. We next describe the use of remote sensing to study flooding. MODIS is the workhorse of remote sensing of flooding; however, a range of other multispectral sensors and synthetic aperture radar (SAR) have also been used for this application. We then describe the use of remote sensing in the study of forest fires – including active fire monitoring, risk area mapping, burn scars estimation, burn severity measurement, fuel load estimation, smog and haze monitoring, and fire management. Finally we describe the use of remote sensing in change-based measurements in which before and after imagery of natural disasters can provide valuable information relating to earthquakes, tsunami, landslides, tornadoes, blowdown/tropical storms, and typhoons/hurricanes. Digital elevation maps (DEMs) are a cross-cutting



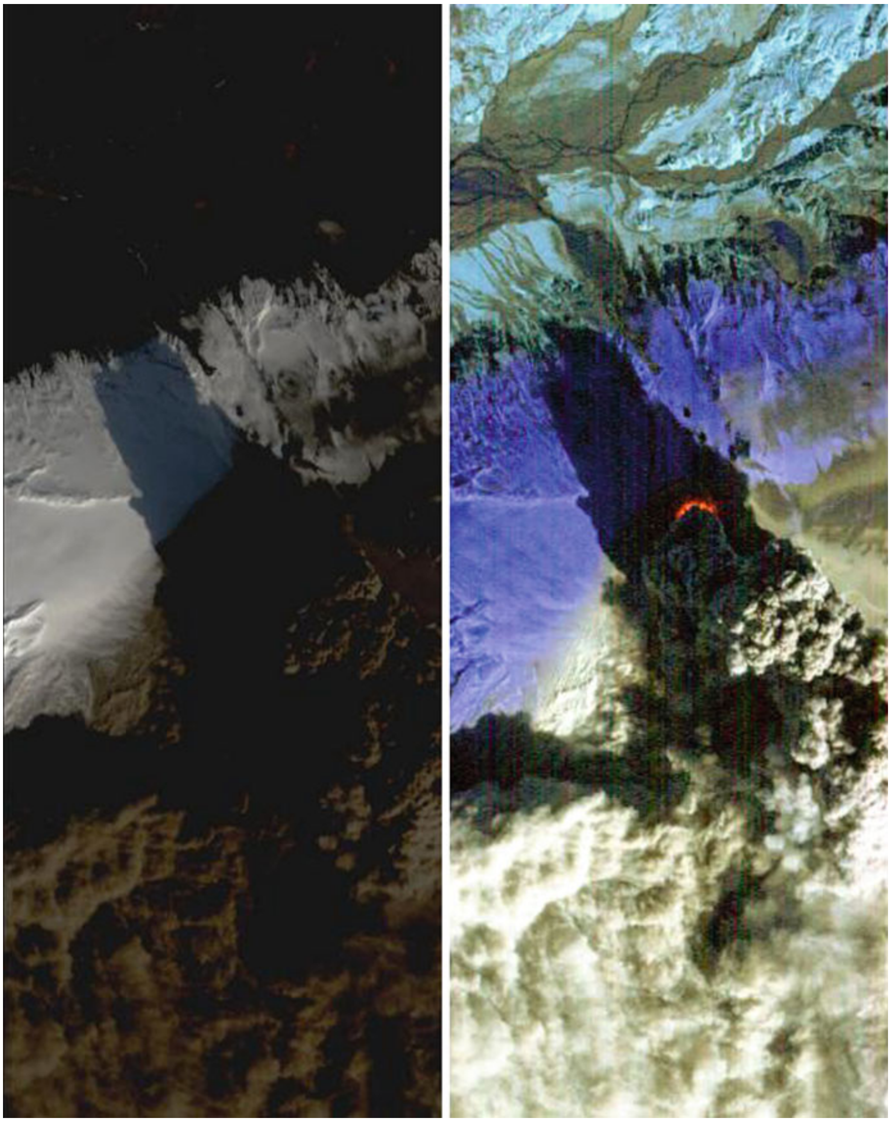
Remote Sensing of Natural Disasters. Figure 1
MODIS imagery of the Eyjafjallajökull eruption (Image courtesy: NASA GSFC/JPL/Caltech)

application of remote sensing. These detailed topographical maps are generally created using remote sensing data and have a wide range of applications relating to natural disasters such as prediction of lava flow movement and lahar flows for volcanoes, flooding, tsunami risk areas, landslide risk areas, fire movement modeling, and others.

Remote Sensing for the Study of Volcanic Activity

Remote sensing has many applications to volcanology. A wide range of instruments have been used to study the thermal emissions of volcanic activity. Measurement of thermal emissions can then drive models to estimate surface lava, effusion rates, and other key physical parameters of volcanic activity. GOES and AVHRR [1] have been used to detect thermal signatures of volcanic activity from space. MODVOLC uses the MODIS sensor (flying on both Terra and Aqua) to detect volcanic thermal signatures [2, 3]. These methods use the multispectral capability of the instrument to fit the thermal emissions to blackbody radiation curves to estimate the surface area emitting at a specific temperature. However, all of these methods are restricted by the low spatial resolution (and corresponding lower thermal sensitivity) of these instruments. While site-specific algorithm parameterization for background thermal signatures can increase sensitivity, higher spatial resolution instruments (typically pointable) can provide more information. In this approach, many other pointable sensors including ASTER [4] and Hyperion [5, 6] have been used to study the thermal emission of volcanic activity.

Remote sensing can also be used to study volcanic plumes. ASTER and MODIS have been used to track such activity [7, 8]. In these applications the plumes and gas emissions of volcanoes can be distinguished from the normal atmosphere due to difference in spectral response from both ash and SO_2 constituents. The AVHRR sensor has also been used in conjunction with ground measurements to study volcanic plumes [9].



Remote Sensing of Natural Disasters. Figure 2
EO-1/ALI true color (*left*) thermal enhanced (*right*) imagery acquired April 17, 2010 (Images courtesy: NASA/JPL/Caltech/EO-1 mission/GSFC/Ashley Davies)

Synthetic aperture radar (SAR) can also be used in an interferometric fashion to study inflation in volcanoes as both a precursor and during eruptions [10].

Tracking of lahar from volcanoes is another application of remote sensing. Visible, Near-Infrared, Synthetic Aperture Radar, and Light Detection And Ranging (LIDAR) sensors have all been used to map lahar flows (see [11] for a comparison of these methods). Note that digital elevation maps, also typically derived by remote sensing, can be used to model (predict) where lahar flows are likely.

The 2010 eruption of the Eyjafjallajökull Volcano in Iceland represents an excellent example of the use of remote sensing in volcanic study. In the weeks immediately following this eruption, data from many space-based sensors, such as MODIS, AIRS, MISR, ASTER, and ALI, were all used in the analysis of the eruption.

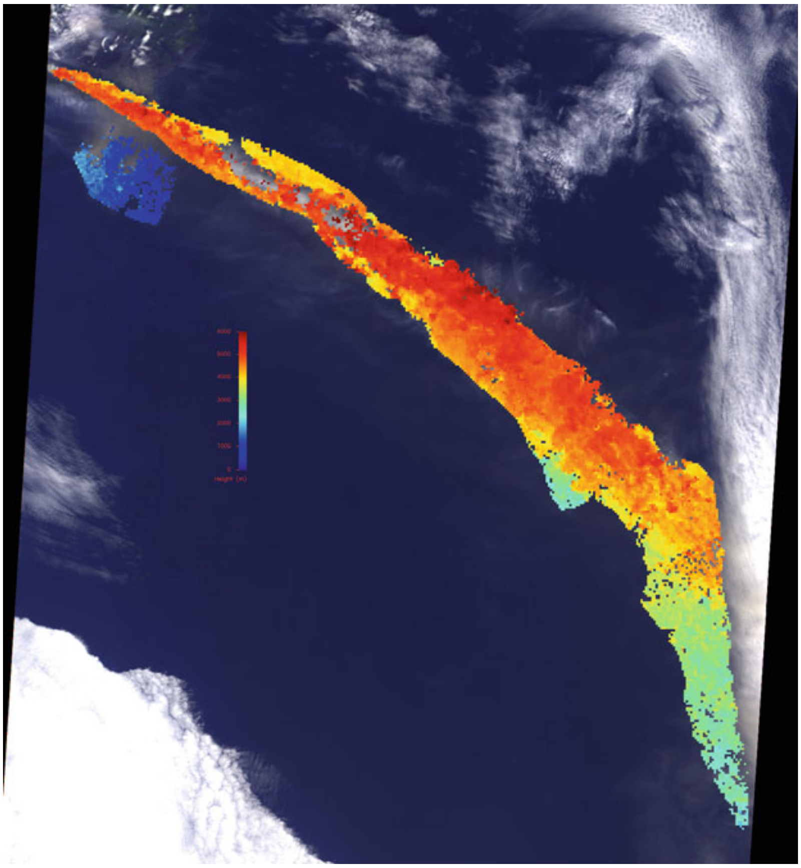
The MODIS instrument provides broad swath (2,000+ km wide) moderate spatial resolution (250–1,000 m/pixel) imagery of Earth in visible to thermal infrared spectra. This workhorse of remote sensing has numerous applications to natural hazards including volcanology. MODIS thermal infrared channels can distinguish ash from other atmospheric conditions as shown in this April 15, 2010, imagery of the Eyjafjallajökull eruption (Fig. 1).

The Advanced Land Imager (ALI) on EO-1 is a targetable instrument sensitive in the visible, short wave infrared, and very near infrared spectra. Figure 2 shows true color (left) and false color to highlight thermal features (right) to highlight the Eyjafjallajökull volcanic activity on April 17, 2010.

The Multi-angle Imaging SpectroRadiometer (MISR) aboard NASA's Terra satellite collected data on ash height when it passed just east of the Eyjafjallajökull Volcano mid-morning on May 7. MISR uses nine different cameras with each camera viewing the event from a different angle to enable hyper-stereo reconstruction of the distance of viewed substances from the instrument. In this application of MISR, the nine angle views are used to reconstruct the distance of the plume from the MISR instrument. Because the altitude of the satellite orbit is known the height of the volcanic plume can then be calculated as shown below in Fig. 3.

Flooding Applications of Remote Sensing

Remote sensing is frequently used in tracking flooding worldwide. The MODIS sensor has been used to study long-term impacts from flooding, with products



Remote Sensing of Natural Disasters. Figure 3
MISR extracted plume height from the Eyjafjallajökull Volcano May 7, 2010 (NASA image courtesy: GSFC/LaRC/JPL/Caltech MISR team)

spanning decades by the Dartmouth Flood Observatory [12, 13] and the University of Maryland [14]. These methods leverage the fact that surface water from flooded areas is very dark and thus can be distinguished spectrally from other substances. For a reliable product cloud shadow areas are removed by looking at multiple images in sequence as MODIS provides at least two daylight overflights per 24 h period. Figure 4 shows a MODIS-derived product developed by the University of Maryland highlighting flooding in Myanmar in 2008.

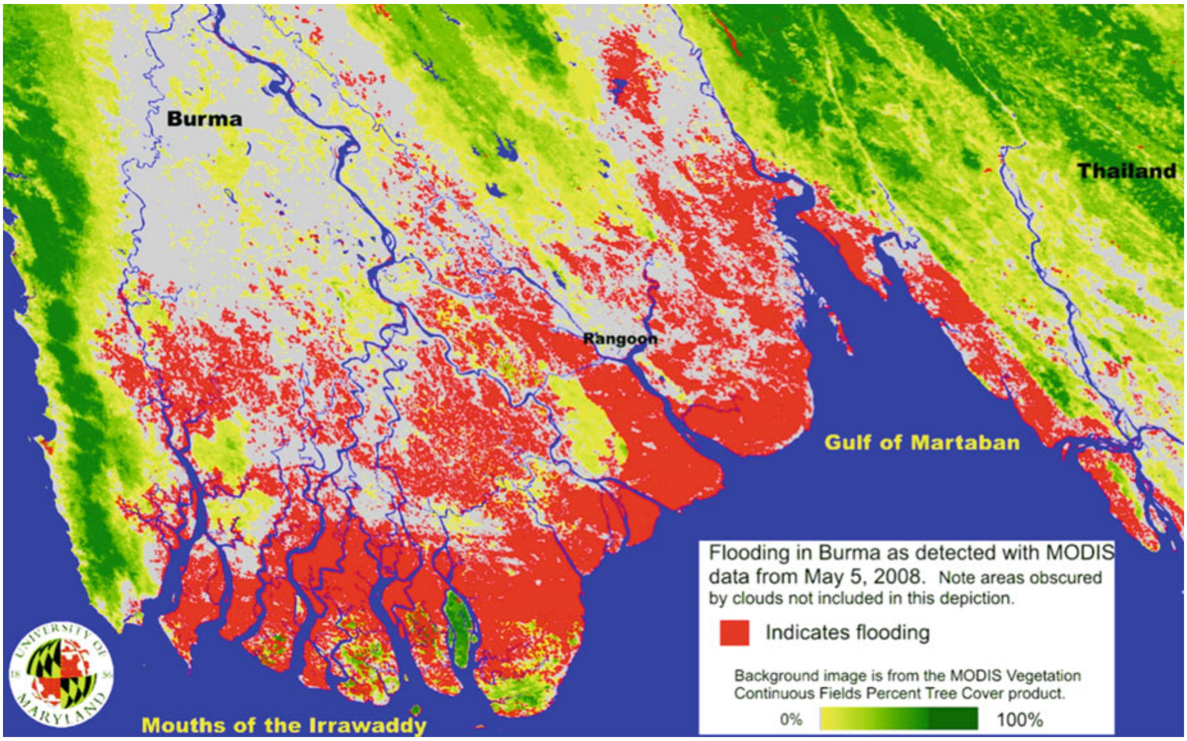
Hyperion [15] and ALI [16] have also been used to study flooding by using spectral analysis methods to automatically derive surface water extent.

Use of multispectral remote sensing to track flooding suffers from the drawback that flooding often occurs in conjunction with cloud cover. Use of radar to track flooding does not suffer from this limitation. QuikSCAT (Quick Scatterometer) has been used to study flooding worldwide [17] providing global coverage maps of flooding. Figure 5 shows QuikSCAT-derived mapping of flooding in China in September 2003.

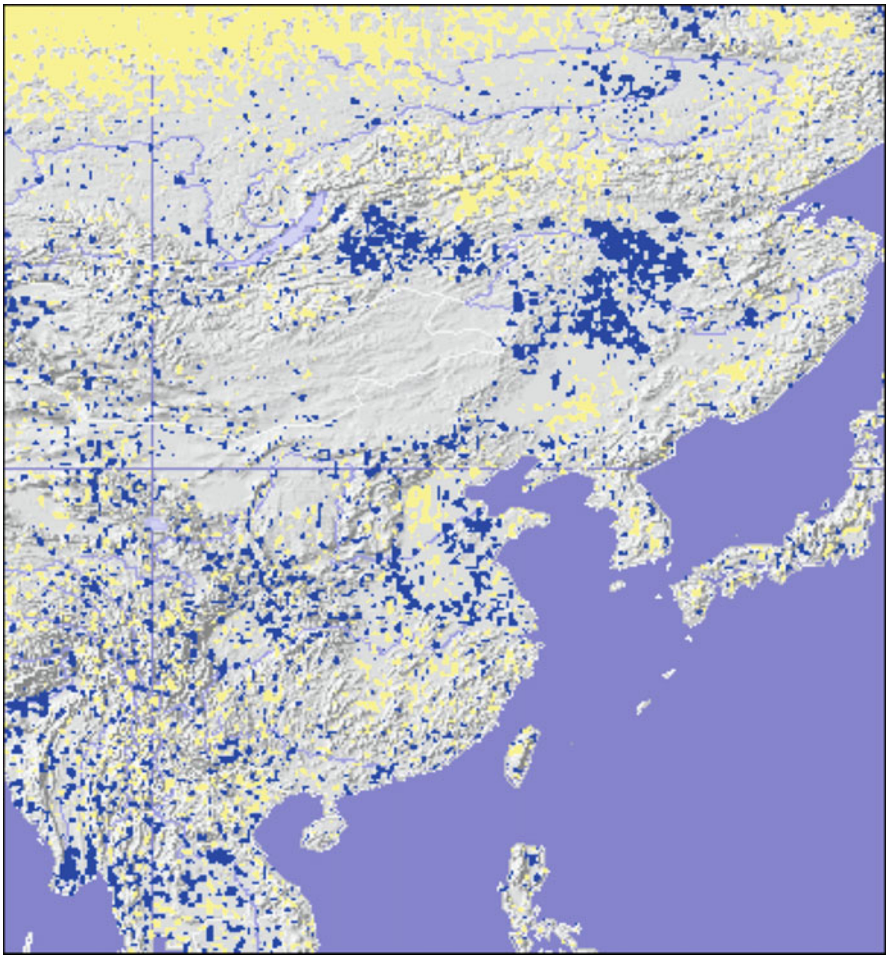
On a regional scale many SARs including Radarsat-1 SAR [18] and ASAR on Envisat [19] have been used in a targeted mode to track flooding. In these applications, changes in the radar returns allow discrimination between flooded areas and other types of land surfaces such as sand, soil, and vegetation. Below Fig. 6a, b shows the formed SAR image and a derived surface water extent classification. The surface water can be distinguished because water has dielectric properties that differ from most land types resulting in a change in the radar returns. However, in cases where weather (e.g., wind) causes a significant change in the water surface roughness it can become more difficult to distinguish between water and land surfaces.

Fire Applications of Remote Sensing

Forest fire is one of the most significant and natural causes of terrestrial biomass change. Forest fires play a major and vital role in the deforestation of the tropical and subtropical regions, especially in the developing and underdeveloped countries. Since it is


Burma
Thailand
Rangoon
Gulf of Martaban
Mouths of the Irrawaddy
Flooding in Burma as detected with MODIS data from May 5, 2008. Note areas obscured by clouds not included in this depiction.
🟥 Indicates flooding
Background image is from the MODIS Vegetation Continuous Fields Percent Tree Cover product.
0%
100%
UNIVERSITY OF MARYLAND

Remote Sensing of Natural Disasters. Figure 4
Flooding in Myanmar May 2008 as derived from MODIS imagery (Images courtesy: University of Maryland)

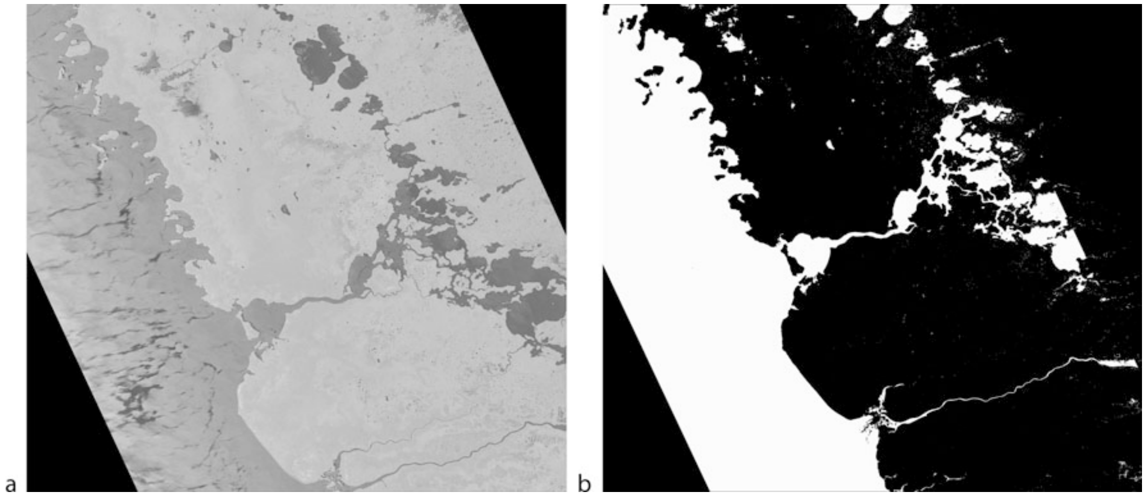


Remote Sensing of Natural Disasters. Figure 5
QuikSCAT-derived flood map for September 2003. Blue indicates flooded areas (Product courtesy: R. Brakenridge/ Dartmouth flood observatory & S. Nghiem/Jet propulsion laboratory)

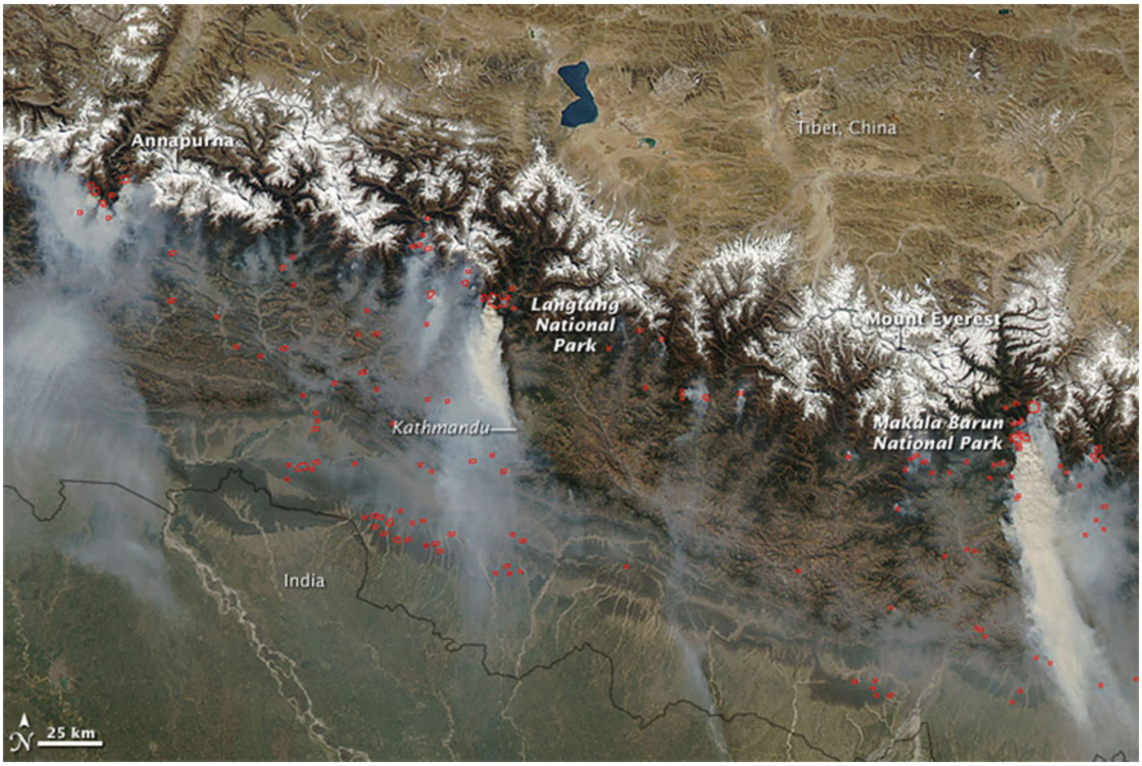
impossible to monitor forest fires of a large area at a ground level, satellite remote sensing is critical. In addition, routine monitoring via satellite remote sensing technology offers the most efficient and cost-effective means for forest fire management over larger areas [20].

An early space system used to study fires is the Advanced Very High Resolution Radiometer (AVHRR) of the National Oceanic and Atmospheric Administration (NOAA) polar orbiting platforms (POES) family. It was originally used for weather monitoring, and then employed as the main sensor in the detection of active fires, or "hotspots," on a global scale and with relatively high temporal frequency. Other sensors include those in the Geostationary Operational Environmental Satellite (GOES) such as the Visible Infrared Spin Scan Radiometer Atmospheric Sounder (VAS) [21] and the GOES Imager [22]. In addition to their main functions, the Defense Meteorological Satellite Program Operational Linescan System (DMSP-OLS) [23, 24] the Along Track Scanning Radiometer (ATSR) [25], the Tropical Rainfall Measuring Mission Visible and Infrared Scanner (TRMM-VIRS) [26] are also used in the forest fire monitoring. However, MODIS is the first sensor specifically designed and developed for the forest fires detection [27, 28].

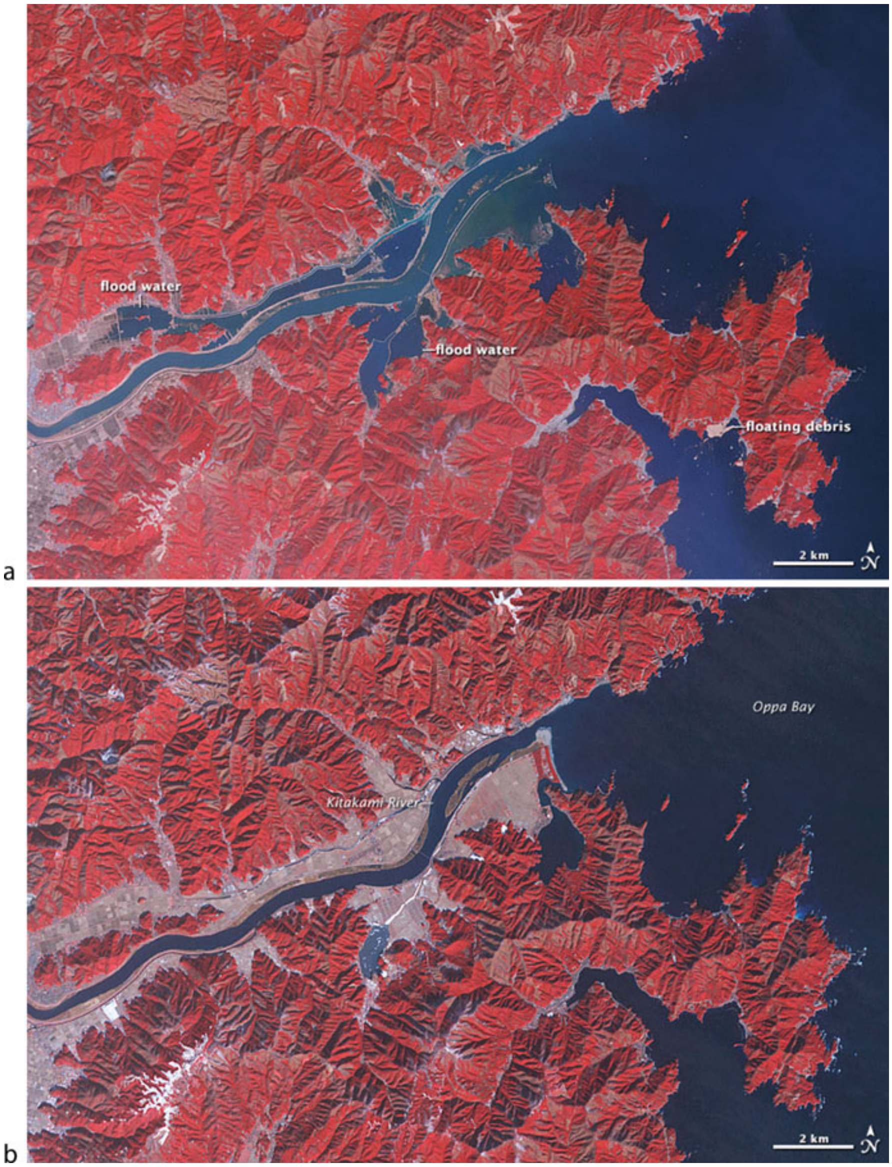
Figure 7 shows MODIS imagery of Nepal acquired in March 2009. In the image active fire detections are



Remote Sensing of Natural Disasters. Figure 6
(**a**) Synthetic aperture radar imagery from the Florida Everglades acquired by the unpiloted aerial vehicle SAR (UAVSAR) L-band SAR in 2009 (Image courtesy: UAVSAR team). (**b**) Surface water classification derived from UAVSAR L-band SAR imagery of Florida Everglades (Image courtesy: UAVSAR team and J. Doubleday/JPL/Caltech)



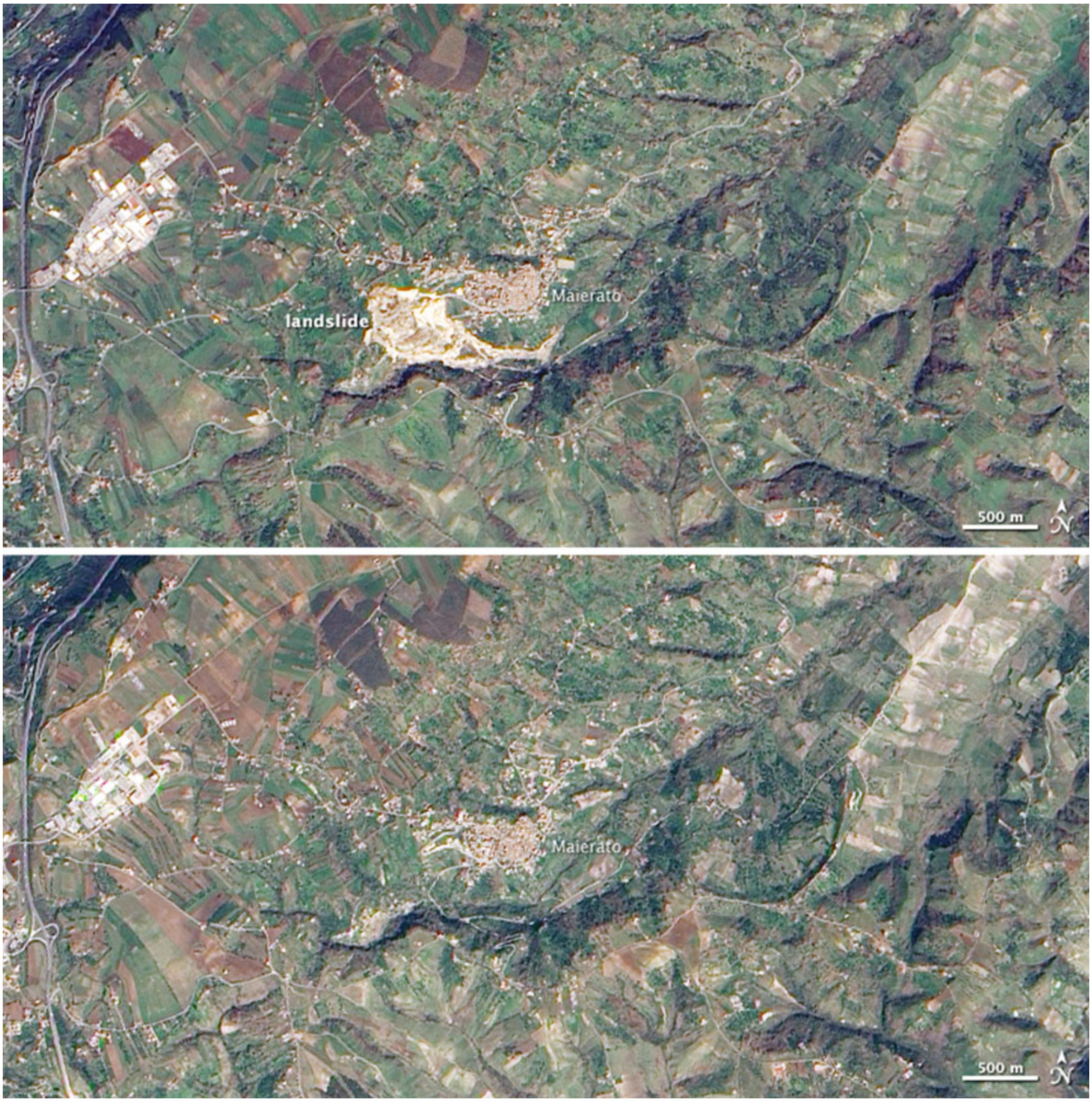
Remote Sensing of Natural Disasters. Figure 7
MODIS imagery showing 2009 fires in Nepal, *red* boxes indicate active fire detections (Image courtesy: NASA and MODIS rapid response team)



Remote Sensing of Natural Disasters. Figure 8
(**a**) ASTER imagery showing the extent of Tsunami-induced flooding at the Kitakami river, March 2011 (NASA earth observatory image by Robert Simmon and Jesse Allen, using data from the GSFC/METI/ERSDAC/JAROS, and USA/Japan ASTER science team). (**b**) ASTER imagery from January 2011 showing pre-tsunami shoreline of the Kitakami river, Japan, January 2011 (NASA earth observatory image by Robert Simmon and Jesse Allen, using data from the GSFC/METI/ERSDAC/JAROS, and USA/Japan ASTER science team)

indicated and the large smoke plumes caused by the fire are quite evident.

An array of technologies used in the satellite remote sensing of forest fires includes spaceborne system/sensors in a variety of tasks which are active fire monitoring [29–32], risk area mapping and assessment [33–36], burn scar estimation and monitoring [37–40], burn severity measurement [41–43], fuel moisture content estimation [44, 45], smog and haze monitoring [46–48], fire behavior [49], and fire management [50].

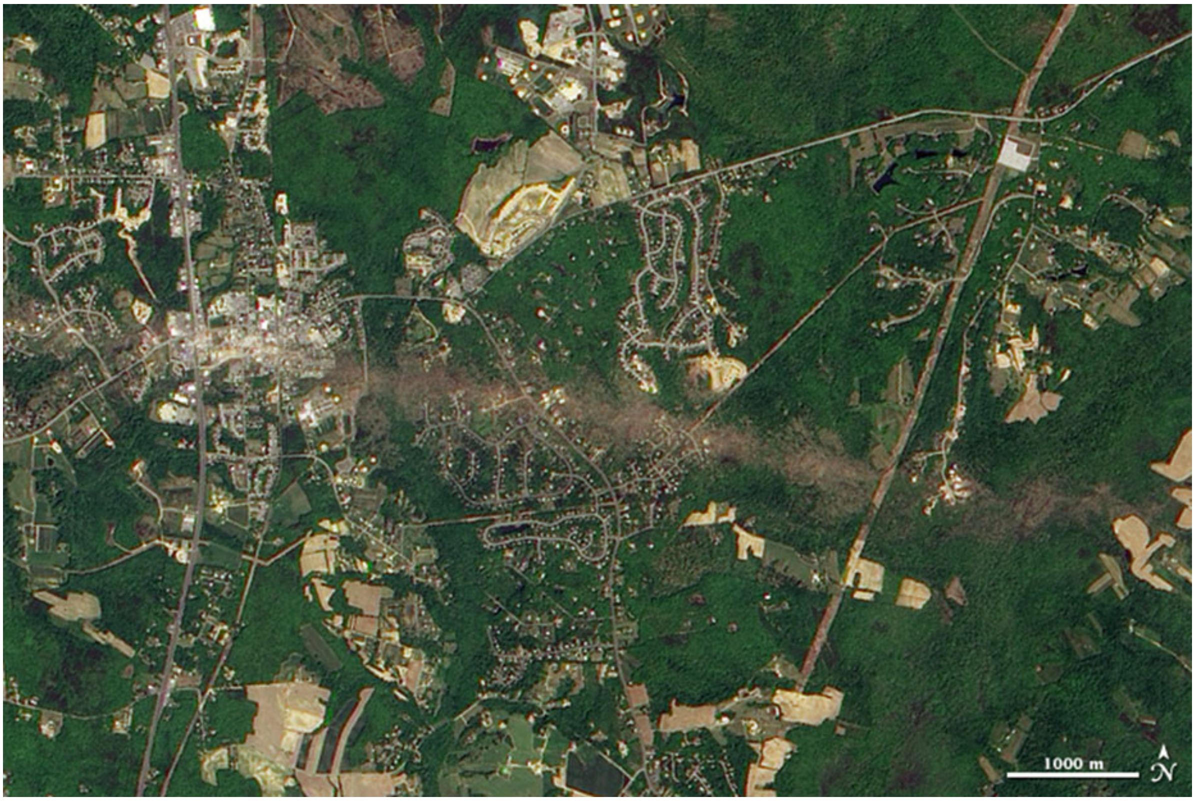


Remote Sensing of Natural Disasters. Figure 9
"After" (*top*) and "before" (*bottom*) imagery of a landslide in Maierato, Italy (Image courtesy of NASA earth observatory and NASA EO-1 team)

One emphasis of current work is to increase the sensitivity of fire detection systems – MODIS capabilities can only detect moderate-sized fires. Another emphasis of current work is to allow for faster delivery of data to enable rapid response to fires before they grow in size and are harder to contain.

Other Applications of Remote Sensing to Natural Disasters

Another major application of remote sensing is the use of "before" and "after" images to assess damage from natural disasters. This type of analysis is relevant to a wide range of natural disasters including tsunamis,



Remote Sensing of Natural Disasters. Figure 10
EO-1 advanced land imager imagery of La Plata, Maryland following tornado in May 2002 (Image courtesy of NASA earth observatory and NASA EO-1 team)

earthquakes, flooding, and tornadoes. For example, Fig. 8 shows the extent of the tsunami-induced flooding in northern Japan following the earthquake–tsunami disaster of March 2011. Figure 8a shows the "after" and Fig. 8b "before." The pair of images clearly indicates the large amount of flood damage and resultant inundation along the Kitakami river.

Another before/after application of satellite imagery for natural disasters is tracking landslide damage. Figure 9 highlights the progression of a landslide in Southern Italy near the town of Maierato using Advanced Land Imager (ALI) data from the EO-1 mission.

Figure 10 shows the devastation caused by a tornado in La Plata, Maryland in April 2002 (imagery from May 2002). In this EO-1 Advanced Land Imager (ALI) imagery, the devastation along the tornado's path is clearly shown in the image as a linear area following the tornadoes path from left to right in the image.

Future Directions

There are many areas of current work aimed at enhancing the utility of remote sensing for assessment, response, and mitigation of natural disasters. Central areas of improvement in remote sensing are directly relevant: improving the resolution of remote sensing, reducing the time lag from acquisition of data to delivery to users, increasing the temporal frequency of coverage, and improving the ability to correct data from degradation due to atmospheric, environmental, or man-made factors.

A major area of improvement is the integration of remote sensing data with environmental models. Increasingly, sensor data are automatically processed and fed into models that perform hindcasting, nowcasting, and forecasting. These models allow for key physical parameters to be estimated even if not directly measureable, for past or ongoing events, as well as the more typical forecasting of how events will progress. These capabilities have dramatic ramifications for natural disasters. For example, such models can be used to estimate whether volcanic activity is increasing or decreasing. Or such capabilities can combine surface water data, topographical and terrain data, and weather forecasts to estimate areas likely at risk for flooding. Active fire mapping, terrain, vegetation, and weather data can be used to estimate likely progressions for active forest fires as a third example. In all of these cases, key data (much of it being remote sensing data) combined with modeling can provide important capabilities for disaster response and mitigation.

An additional exciting new area is the use of "sensorwebs" for environmental monitoring and specifically in the use of sensing for natural disasters. In a sensorweb, data are processed "on the fly" and used to actively reconfigure other portions of the sensorweb to better acquire data and track an evolving phenomena [51]. For example, detections of volcanic activity by one satellite might trigger observations by other satellites [52]. Or weather reports in concert with satellite rainfall observations might trigger observations of likely flood areas [16]. These techniques, in concert with modeling, will enhance future sensing of natural disasters.

Conclusions

Remote sensing has a wide range of applications to risk assessment, tracking, response, and mitigation of natural disasters. In this entry, we have outlined only a few of the applications to several categories of natural disasters focusing volcanoes, flooding, and fires. Remote sensing also provides secondary products such as topographical maps that are widely used in modeling natural disaster phenomena. Topographical maps can be used to estimate risk (to tsunamis, flooding, landslide, etc.) as well as provide inputs to model and predict flood progression, and lava and lahar flows. All of these areas and many more are the subject of active research with new applications and techniques being developed continuously.

Acknowledgments

Portions of the research described in this entry were carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

Bibliography

1. Harris AJ, Flynn LP, Dean K, Wooster EM, Okubo C, Mouginis-Mark P et al (2000) Real-time satellite monitoring of volcanic hot spots. Geophys Monogr 116:139–159
2. Wright R, Flynn LP, Garbeil H, Harris A, Piler E (2003) Automated volcanic eruption detection using MODIS. Remote Sens Environ 82:135–155
3. Wright R, Flynn LP, Garbeil H, Harris A, Piler E (2004) MODVOLC: near-real-time thermal monitoring of global volcanism. J Volcanol Geoth Res 135:29–49
4. Pieri D, Abrams M (2004) ASTER watches the world's volcanoes: a new paradigm for volcanological observations from orbit. J Volcanol Geoth Res 135(1–2):13–28
5. Davies AG, Chien S, Baker V, Doggett T, Dohm J, Greeley R, Ip F, Castano R, Cichy B, Rabideau G, Tran D, Sherwood R (2006) Monitoring active volcanism with the autonomous sciencecraft, experiment on EO-1. Remote Sens Environ 101:427–446
6. Davies AG, Calkins J, Scharenbroich L, Vaughan G, Wright R, Kyle P, Castano R, Chien S, Tran D (2008) Multi-instrument remote and in situ observations of the erebus volcano (Antarctica) lava lake in 2005: a comparison with the Pele lava lake on the Jovian moon Io. J Volcanol Geoth Res 177(3): 705–724
7. Pieri D, Gubbels T, Hufford G, Olsson G, Realmuto V (2006) Assessing mesoscale volcanic aviation hazards using ASTER. Eos Trans AGU 87(52). Fall meet. Suppl, Abstract H33E-1554
8. Novak MA, Watson IM, Delgado-Granados H, Rose WI, Cardenas-Gonzalez L, Realmuto VJ (2008) Volcanic emissions from Popocatépetl volcano, Mexico, quantified using moderate resolution imaging spectroradiometer (MODIS) infrared data: a case study of the December 2000–January 2001 emissions. J Volcanol Geoth Res 170:76–85
9. Andronico D, Spinetti C, Cristaldi A, Buongiorno MF (2009) Observations of Mt. Etna volcanic ash plumes in 2006: an integrated approach from ground-based and polar satellite NOAA–AVHRR monitoring system. J Volcanol Geoth Res 180(2–4):135–147
10. Tralli DM, Blom RG, Zlotnicki V, Donnellan A, Evans DL (2005) Satellite remote sensing of earthquake, volcano, flood, landslide and coastal inundation hazards. ISPRS J Photogramm 59(4):185–198
11. Joyce K, Samsonov S, Manville V, Jongens R, Graettiner A, Cronin S (2009). Remote sensing data types and techniques for lahar path detection: a case study at Mt Ruapehu, New Zealand. Remote Sens Environ113(8) 1778–1786; Wright R, Flynn L, Garbeil H, Harris A, Piler E (2002). Automated volcanic eruption detection using MODIS. Remote Sens Environ 82(1) 135–155
12. Brakenridge GR, Anderson E (2005) MODIS-based flood detection, mapping, and measurement: the potential for operational hydrological applications. In: Transboundary floods, proceeding of NATO advanced research workshop, Baile Felix – Oradea, 4–8 May 2005

13. Brakenridge G R (2010). DFO MODIS flood retrieval algorithm. http://floodobservatory.colorado.edu/Tech.html. Accessed 21 Mar 2011
14. Carroll M, Townshend J, Noojipady P, DiMiceli C, Sohlberg R (2009). Surface water dynamics derived from the MODIS data record. Spring AGU
15. Ip F, Dohm J, Baker V, Doggett T, Davies A, Castano R, Chien S, Cichy B, Greeley R, Sherwood R, Tran D, Rabideau G (2006) Flood detection and monitoring with the autonomous sciencecraft experiment onboard EO-1. Remote Sens Environ 101(4):463–481
16. Chien S, Doubleday J, Mclaren D, Tran D, Khunboa C, Leelapatra W, Plergamon V, Tanpipat V, Raghavendra C, Mandl D (2011). Using multiple space assets with In-situ measurements to track flooding in Thailand. In: 34th international symposium on remote sensing of environment, Sydney
17. Brakenridge GR, Nghiem SV, Anderson E, Chien S (2005) Space-based measurement of river runoff. Eos Trans AGU 86(19):185. doi:10.1029/2005EO190001
18. Townsend P (2001) Mapping seasonal flooding in forested wetlands using multi-temporal Radar- sat SAR. Photogramm Eng Remote Sens 67:3055–3074
19. Marti-Cardona B, Lopez-Martinez C, Dolz-Ripolles J, Blade-Castellet E (2010) ASAR polarimetric, multi-incidence angle and multitemporal characterization of Doñana wetlands for flood extent monitoring. Remote Sens Environ 114(11):2802–2815
20. Justice CO, Malingreau JP, Seltzer A (1993) Satellite remote sensing of fires: potential and limitations. In: Crutzen P, Goldammer J (eds) Fire in the environment: the ecological, atmospheric, and climatic importance of vegetation fires. Wiley, New York, pp 77–88
21. Prins EM, Menzel WP (1992) Geostationary satellite detection of biomass burning in South America. Int Remote Sens 13:2783–2799
22. Menzel WP, Prins EM (1996). Monitoring fire activity in western hemisphere with the new generation of geostationary satellites. In: 22nd conference on agricultural and forest meteorology with symposium on fire and forest meteorology, Atlanta, 28 Jan−2 Feb, 1996, pp 272–275
23. Elvidge C, Kroehl HW, Kihn EA, Baugh KE, Davis ER, Hao WM (1996) Algorithm for the retrieval of fire pixels from DMSP operational linescan system data. In: Levine JS (ed) Biomass burning and global change, vol 1. The MIT Press, Cambridge, MA, pp 73–85
24. Elvidge C, Dee WP, Elaine P, Eric AK, Jackie K, Kimberly EB (1998) Remote sensing change detection: environmental monitoring methods and applications, wildfire detection with meteorological satellite data: results from New Mexico during June of 1996 using GOES, AVHRR, and DMSP-OLS. CRC Press, Boca Raton, FL, pp 103–121
25. Arino O, Rosaz J (1999). 1997 and 1998 world ASTR fire atlas using ERS-2 ATSR-2 data. In: Neuenschwander LF, Tyan KC, Golberg GE (eds) Proceedings of the joint fire science conference, Boise, Idaho, 15–17 June 1999. University of Idaho and the International Association of Wildland Fire, Boise, pp 177–182
26. Giglio L, Kendall JD, Tucker CJ (2000) Remote Sensing of fires with TRMM VIRS. Int J Remote Sens 21:203–207
27. Giglio L, Descloitres J, Justice CO, Kaufman YJ (2003) An enhanced contextual fire detection algorithms for MODIS. Remote Sens Environ 87:273–282
28. Justice CO, Giglio L, Korontzi S, Owens J, Morisette JT, Roy D, Descloitres J, Alleaume S, Petitcolin F, Kaufman Y (2002) The MODIS fire products. Remote Sens Environ 83:244–262
29. Roy DP, Boschetti L, Justice CO, Ju J (2008) The collection 5 MODIS burned area product – Global evaluation by comparison with the MODIS active fire product. Remote Sens Environ 112(9):3690–3707. doi:10.1016/j.rse.2008.05.013, ISSN 0034-4257
30. Ressl R, Lopez G, Cruz I, Colditz RR, Schmidt M, Ressl S, Jimenez R (2009) Operational active fire mapping and burnt area identification applicable to Mexican nature protection areas using MODIS and NOAA-AVHRR direct readout data. Remote Sens Environ 113(6):1113–1126. doi:10.1016/j.rse.2008.10.016, ISSN 0034-4257
31. Amraoui M, DaCamara CC, Pereira JMC (2010) Detection and monitoring of African vegetation fires using MSG-SEVIRI imagery. Remote Sens Environ 114(5):1038–1052. doi:10.1016/j.rse.2009.12.019, ISSN 0034-4257
32. Tanpipat V, Honda K, Nuchaiya P (2009) MODIS hotspot validation over Thailand. Remote Sens 1:1043–1054. doi:10.3390/rs1041043, ISSN 2072-4292
33. Chuvieco E, Aguado I, Yebra M, Nieto H, Salas J, Martin MP, Vilar L, Martinez J, Martin S, Ibarra P, Riva JDL, Baeza J, Rodriguez F, Molina JR, Herrera MA, Zamora R (2010). Development of a framework for fire risk assessment using remote sensing and geographic information system technologies. Ecol Model 221(1 Special issue on spatial and temporal patterns of wildfires: models, theory, and reality) 46–58, doi: 10.1016/j.ecolmodel.2008.11.017, ISSN 0304–3800
34. Maeda EE, Arcoverde GFB, Pellikka PKE, Shimabukuro YE (2011) Fire risk assessment in the Brazilian Amazon using MODIS imagery and change vector analysis. Applied Geography 31(1):76–84. doi:10.1016/j.apgeog.2010.02.004, ISSN 0143-6228
35. Schneider P, Roberts DA, Kyriakidis PC (2008) A VARI-based relative greenness from MODIS data for computing the Fire Potential Index. Remote Sens Environ 112(3):1151–1167. doi:10.1016/j.rse.2007.07.010, ISSN 0034-4257
36. Huesca M, Litago J, Palacios-Orueta A, Montes F, Sebastian-Lopez A, Escribano P(2009). Assessment of forest fire seasonality using MODIS fire potential: a time series approach. Agr Forest Meteorol 149(11 Special section on water and carbon dynamics in selected ecosystems in China) 1946–1955, doi: 10.1016/j.agrformet.2009.06.022, ISSN 0168–1923
37. Giglio L, Loboda T, Roy DP, Quayle B, Justice CO (2009) An active-fire based burned area mapping algorithm for the MODIS sensor. Remote Sens Environ 113(2):408–420. doi:10.1016/j.rse.2008.10.006, ISSN 0034-4257

38. Libonati R, DaCamara CC, Pereira JMC, Peres LF (2010) Retrieving middle-infrared reflectance for burned area mapping in tropical environments using MODIS. Remote Sens Environ 114(4):831–843. doi:10.1016/j.rse.2009.11.018, ISSN 0034-4257
39. Dubinin M, Potapov P, Lushchekina A, Radeloff VC (2010) Reconstructing long time series of burned areas in arid grasslands of southern Russia by satellite remote sensing. Remote Sens Environ 114(8):1638–1648. doi:10.1016/j.rse.2010.02.010, ISSN 0034-4257
40. Zhang X, Kondragunta S (2008) Temporal and spatial variability in biomass burned areas across the USA derived from the GOES fire product. Remote Sens Environ 112(6):2886–2897. doi:10.1016/j.rse.2008.02.006, ISSN 0034-4257
41. Veraverbeke S, Lhermitte S, Verstraeten WW, Goossens R (2011) A time-integrated MODIS burn severity assessment using the multi-temporal differenced normalized burn ratio (dNBRMT). Int J Appl Earth Obs Geoinformation 13(1):52–58. doi:10.1016/j.jag.2010.06.006, ISSN 0303–2434
42. Santis AD, Asner GP, Vaughan PJ, Knapp DE (2010) Mapping burn severity and burning efficiency in California using simulation models and landsat imagery. Remote Sens Environ 114(7):1535–1545. doi:10.1016/j.rse.2010.02.008, ISSN 0034–4257
43. Fox DM, Maselli F, Carrega P (2008) Using SPOT images and field sampling to map burn severity and vegetation factors affecting post forest fire erosion risk. CATENA 75(3):326–335. doi:10.1016/j.catena.2008.08.001, ISSN 0341–8162
44. Yebra M, Chuvieco E, Riano D (2008) Estimation of live fuel moisture content from MODIS images for fire risk assessment. Agr Forest Meteorol 148(4):523–536. doi:10.1016/j.agrformet.2007.12.005, ISSN 0168–1923
45. Peterson HS, Roberts DA, Dennison PE (2008) Mapping live fuel moisture with MODIS data: A multiple regression approach. Remote Sens Environ 112(12):4272–4284. doi:10.1016/j.rse.2008.07.012, ISSN 0034–4257
46. Hyer EJ, Chew BN (2010) Aerosol transport model evaluation of an extreme smoke episode in Southeast Asia. Atmos Environ 44(11):1422–1427. doi:10.1016/j.atmosenv.2010.01.043, ISSN 1352–2310
47. Henderson SB, Burkholder B, Jackson PL, Brauer M, Ichoku C (2008) Use of MODIS products to simplify and evaluate a forest fire plume dispersion model for PM10 exposure assessment. Atmos Environ 42(36):8524–8532. doi:10.1016/j.atmosenv.2008.05.008, ISSN 1352–2310
48. Zhang X, Kondragunta S, Schmidt C, Kogan F (2008) Near real time monitoring of biomass burning particulate emissions (PM2.5) across contiguous United States using multiple satellite instruments. Atmos Environ 42(29):6959–6972. doi:10.1016/j.atmosenv.2008.04.060, ISSN 1352–2310
49. Balch JK, Nepstad DC, Curran LM, Brando PM, Portela O, Guilherme P, Reuning-Scherer JD Jr, Carvalho OD (2011) Size, species, and fire behavior predict tree and liana mortality from experimental burns in the Brazilian Amazon. Forest Ecol Manage 261(1):68–77. doi:10.1016/j.foreco.2010.09.029, ISSN 0378–1127
50. Bonazountas M, Kallidromitou D, Kassomenos P, Passas N (2006) A decision support system for managing forest fire casualities. Environ Manage. doi:10.1016/j.jenvman.2006.06.016
51. Chien S, Cichy B, Davies A, Tran D, Rabideau G, Castano R, Sherwood R, Mandl D, Frye S, Shulman S, Jones J, Grosvenor S (2005) An autonomous earth observing sensorweb. IEEE Intell Syst 20:16–24
52. Chien S, Davies A, Doubleday J, Tran D, Jones S, Kjartansson E, Vogfjord K, Gudmundsson M, Thordarson T, Mandl D (2011). Integrating multiple space and ground sensors to track volcanic activity. In: 34th international symposium on remote sensing of environment, Sydney

Remote Sensing of Ocean Color

Heidi M. Dierssen, Kaylan Randolph
Department of Marine Sciences, University of Connecticut, Groton, CT, USA

Article Outline

Glossary
Definition of the Subject, Relevance, Motivation
Introduction
Optical Properties of the Water Column
Basics of Ocean Color Remote Sensing
Ocean Color Algorithms
Applications for Oceanography
Applications for Environmental Monitoring
Future Directions
Bibliography

Glossary

Absorption, $\boldsymbol{a(\lambda)}$ The fraction of a collimated beam of photons in a particular wavelength (λ), which is absorbed or scattered per unit distance within the medium (units 1/length or m^{-1}). Photons which are absorbed by ocean water alter the spectral distribution of light that can be observed remotely.

Apparent optical properties (AOP) Optical properties which depend primarily on the medium itself but have a small dependence on the ambient light field. Typically, AOPs are derived from measurements of the ambient light field, particularly upwelling and downwelling radiance and irradiance. Principal AOPs include irradiance reflectance, remote sensing reflectance, and the diffuse attenuation coefficients.

Backscattering, $b_b(\lambda)$ Light of a particular wavelength (λ) that is scattered in a direction 90–180° away from its original path (i.e., backward hemisphere). Backscattered light is what is measured as ocean color in remote sensing, namely, downward propagating sunlight that has been redirected back toward the sea surface and out into the atmosphere. For natural waters, only a few percent of the light entering the ocean is backscattered out.

Colored or chromophoric dissolved organic material (CDOM) CDOM is yellow-brown in color and absorbs primarily ultraviolet and blue light decreasing exponentially with increasing wavelength. Produced from the decay of plant material, it consists mainly of humic and fulvic acids and is operationally defined as substances that pass though a 0.2 μm filter.

Diffraction Light which propagates or bends along the boundary of two different mediums with different indices of refraction.

Diffuse attenuation coefficient, K(λ) A normalized depth derivative that describes the rate of change of light, plane incident irradiance, with depth. Sunlight underwater typically decreases exponentially with depth.

Index of refraction (real), *n* The speed of light in a medium, c_{med}, relative to the speed of light in a vacuum, c_v expressed as $n = c_v/c_{med}$. The real index of refraction determines the scattering of light at the boundary between two different mediums and within the medium from thermal and molecular fluctuations. The relative refractive index, n', is the ratio of the speed of light within the medium, c_m, to the speed of light within a particle, c_p. As n' deviates from 1, the scattering caused by the particle increases for a general size and shape particle (e.g., minerals and bubbles).

Inherent optical properties (IOP) Optical properties which depend on the medium itself and are independent of the ambient light field. IOPs are defined from a parallel beam of light incident on a thin layer of the medium. Two fundamental IOPs are the absorption (a) and the volume scattering coefficient (β), which describe how light is either absorbed or directionally scattered by ocean water.

Irradiance (downward planar), $E_d(\lambda)$ The incremental amount of radiant energy per unit time (W) incident on the sensor area (m^{-2}) from all solid angles contained in the upper hemisphere, expressed per unit wavelength of light (λ, nm^{-1}). This is used to measure the amount of spectral energy from the sun reaching the sea surface.

Irradiance reflectance, R(λ) The ratio of the upwelling irradiance, $E_u(\lambda)$, to the plane downwelling irradiance, $E_d(\lambda)$, in different wavelengths (λ).

Optical depth, ζ A measure of how opaque a medium is to radiation. The optical depth is a function of the geometric depth and the vertical attenuation coefficient.

Optically shallow waters An aquatic system where the spectral reflectance off the bottom contributes to radiance measured above the sea surface and is defined by the water clarity, bottom depth, and bottom composition.

Photosynthetically available radiation (PAR) The integrated photon flux (photons per second per square meter) within the 400–700 nm wavelength range at the ocean surface. PAR is the total energy available to phytoplankton for photosynthesis and is reported in units of Q m^{-2} s^{-1}, where Q is quanta, or in μE m^{-2} s^{-1}, where E is Einsteins.

Radiance, $L(\lambda)$ The incremental amount of radiant energy per unit time (in Watts) incident on the sensor area (m^{-2}) in a solid angle view (sr^{-1}) per unit wavelength (λ) of light (nm^{-1}). A satellite measures radiance.

Reflection At the boundary of two different mediums with different indices of refraction, a certain amount of radiation is returned at an angle equal to the angle of incidence.

Refraction The direction of light propagation changes, or is bent, at the boundary between two mediums with different indices of refraction. The refracted light bends toward the normal boundary when the index of refraction increases from one medium to another and away from the normal boundary when the index of refraction decreases from one medium to another.

Remote sensing reflectance, $R_{rs}(\lambda)$ A specialized ratio used for remote sensing purposes formulated as the ratio of the spectral water-leaving radiance, $L_w(\lambda)$, to the plane irradiance incident on the water, $E_d(\lambda)$.

It represents the spectral distribution of sunlight penetrating the sea surface that is backscattered out again and potentially measured remotely. Theoretically, it is proportional to spectral backscattering $b_b(\lambda)$ and inversely proportional to absorption $a(\lambda)$ of the surface water column.

Water-leaving radiance, $L_w(\lambda)$ The component of the radiance signal measured above the water consisting of photons that have penetrated the water column and been backscattered out through the air-sea interface. It does not include photons reflected off the sea surface, also called sun glint.

Definition of the Subject, Relevance, Motivation

The oceans cover over 70% of the earth's surface and the life inhabiting the oceans play an important role in shaping the earth's climate. Phytoplankton, the microscopic organisms in the surface ocean, are responsible for half of the photosynthesis on the planet. These organisms at the base of the food web take up light and carbon dioxide and fix carbon into biological structures releasing oxygen. Estimating the amount of microscopic phytoplankton and their associated primary productivity over the vast expanses of the ocean is extremely challenging from ships. However, as phytoplankton take up light for photosynthesis, they change the color of the surface ocean from blue to green. Such shifts in ocean color can be measured from sensors placed high above the sea on satellites or aircraft and is called "ocean color remote sensing." In open ocean waters, the ocean color is predominantly driven by the phytoplankton concentration and ocean color remote sensing has been used to estimate the amount of chlorophyll *a*, the primary light-absorbing pigment in all phytoplankton. For the last few decades, satellite data have been used to estimate large-scale patterns of chlorophyll and to model primary productivity across the global ocean from daily to interannual timescales. Such global estimates of chlorophyll and primary productivity have been integrated into climate models and illustrate the important feedbacks between ocean life and global climate processes. In coastal and estuarine systems, ocean color is significantly influenced by other light-absorbing and light-scattering components besides phytoplankton. New approaches have been developed to evaluate the ocean color in relationship to colored dissolved organic matter, suspended sediments, and even to characterize the bathymetry and composition of the seafloor in optically shallow waters. Ocean color measurements are increasingly being used for environmental monitoring of harmful algal blooms, critical coastal habitats (e.g., seagrasses, kelps), eutrophication processes, oil spills, and a variety of hazards in the coastal zone.

Introduction

Remote sensing of ocean color allows for the estimation of phytoplankton biomass and carbon fixation over the global ocean. From these data, approximately half of the global carbon fixation is estimated to occur by ocean phytoplankton, accounting for roughly 50 Gt C year^{-1} [1, 2]. Phytoplankton are the base of the marine food web, responsible for producing organic carbon from carbon dioxide. The premise behind ocean color remote sensing is to relate the intensity and spectral distribution of visible light reflected out of the water ("ocean color") to the biological and biogeochemical processes that influence the optical properties of the water column ("bio-optical properties") [3]. The distribution, abundance, and temporal variation in various biological, physical, and chemical processes can be observed synoptically from local and regional to global spatial scales from sensors placed on satellites or aircraft. Ocean color remote sensing provides the long-term, continuous time series of phytoplankton biomass and productivity data necessary for global carbon cycle and climate research [4–6], but the uses of ocean color data are increasingly diverse from military to environmental monitoring applications [7].

Phytoplankton have a marked influence on the subsurface and emergent light field [8]. The light harvesting systems of phytoplankton, including the chlorophyll *a* pigment which is ubiquitous among phytoplankton species, absorb light across the visible spectrum and influence the color of the near-surface ocean [9]. An increase in absorption, or reduction in reflectance, in the blue relative to the green portion of the spectrum can be empirically related to chlorophyll *a* concentration [10]. In other words, as phytoplankton are added to the water column, more blue light is absorbed and the reflected color changes from blue to green. The advent of space-based ocean color sensors in

1978 with NASA's Coastal Zone Color Scanner (CZCS) and the follow on Sea-viewing Wide Field of View Sensor (SeaWiFS) in 1997 greatly enhanced the understanding of phytoplankton distribution and concentration in the ocean [11]. Satellite ocean color imagery provides estimates of phytoplankton abundance across all ocean basins (Atlantic, Pacific, Indian, Arctic, and Southern Oceans) and quantifies the variability from seasonal to interannual timescales.

Over the last several decades, ocean color has expanded beyond chlorophyll and a whole field has emerged to study how the nature of the upwelling light field changes as a function of the quantity and composition of a variety of constituents in the near-surface ocean, including biogenic and nonbiogenic inorganic material, nonliving and living organic material (i.e., phytoplankton, bacteria and viruses), dissolved substances, and benthic habitats. Ocean color research has sought to define the fundamental relationship between the inherent optical properties of the ocean, or the absorption and scattering properties of the constituents, and water-leaving radiance. With improved technology, including radiometers with better spectral resolution, calibration, and a high signal-to-noise ratio, and in situ optical instrumentation, which provided a description of the optical properties of oceanic constituents, biogeochemical parameters are being estimated with greater accuracy and precision. Ocean color remote sensing has moved beyond estimations of chlorophyll alone and is now used to measure total suspended sediment, colored dissolved organic material, particulate inorganic carbon, and phytoplankton functional groups, as well as critical habitats and hazards influencing pelagic and coastal waters.

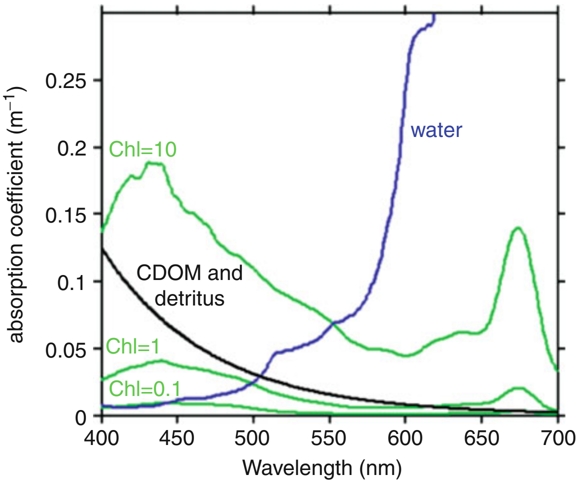
Optical Properties of the Water Column

Scattering and absorption of photons, the basic unit of light energy, in the surface ocean determines the intensity and spectral shape of the water-leaving light signal measured at an ocean color sensor. Photons that propagate into the ocean interact with water molecules, dissolved and particulate matter and are either absorbed or scattered. Because most of the light is propagated downward into the water column, only a small amount of the signal is scattered back out of the water column and measured remotely. The bulk optical properties of water are used to describe how the spectral and directional distribution of photons is altered within the natural water body.

Inherent Optical Properties

The absorption and scattering properties of water molecules and the dissolved and particulate constituents within the water are called inherent optical properties (IOPs). IOPs do not depend on the ambient light conditions, but are a function of the medium alone. The two IOPs commonly used for remote sensing purposes include the *absorption* (a) and scattering (b) coefficients, which refer to the fraction of incident light, a single, narrow, collimated beam of photons, which is absorbed or scattered per unit distance within the medium (units 1/length or m^{-1}). The scattering coefficient stems from the volume scattering function (β), which is the differential scattering cross section per unit volume per solid angle, and is calculated as the integral over all directions (0–180°). The attenuation coefficient (c) accounts for the reduction in light intensity due to absorption and scattering processes combined.

Both absorption and scattering processes can change the color of the ocean as observed from a satellite. Oceanic constituents that are primarily responsible for absorption of photons include water molecules, phytoplankton pigments, particulate detritus, and *colored or chromophoric dissolved organic material* (CDOM) (Fig. 1). Pure water is increasingly effective at absorbing light at wavelengths greater than 550 nm and absorbs minimally in the blue and green portion of the visible spectrum. Conversely, CDOM, operationally defined as all of the colored material that passes through a 0.2 μm filter, absorbs maximally in the ultraviolet and blue portion of the spectrum, decreasing exponentially with wavelength at a rate which is related to the composition, or degradation state, of the material. CDOM is generally comprised of humic and fulvic acids and small colloidal material released through the degradation of plant tissue, whether in soils or in water [12, 13]. Commonly, CDOM is modeled with an exponential function, but a hyperbolic model may be more accurate [14]. Nonliving particulate material, called detritus or tripton, absorbs in a manner similar to CDOM and the two components are difficult to differentiate spectrally.



Remote Sensing of Ocean Color. Figure 1 Absorption spectrum for different constituents in seawater including water molecules, chromophoric dissolved organic matter and detritus, and phytoplankton contributions bio-optically modeled for chlorophyll at 0.1, 1 and 10 mg m^{-3} [16]

Phytoplankton absorb light in a complex manner related to the composition and quantity of their photosynthetic pigments, molecules structured to absorb photons within the visible range of 400–700 nm, dubbed *photosynthetically available radiation* or PAR. There are three distinct classes of pigments, namely, chlorophylls, carotenoids, and biliproteins [101]. All phytoplankton contain chlorophyll *a* and most contain chlorophylls *b* and/or *c*. Chlorophylls *a*, *b*, and *c* have two strong absorption bands in the red and blue portions of the spectrum. Chlorophyll *a* absorption is low in the green (450–650 nm) portion of the spectrum. The presence of chlorophylls *b* and *c* extend the range of light available for photosynthesis further into both the short- and regions. Carotenoids, of which there are many types, have both light harvesting and photoprotective functions. Finally, some phytoplankton contain red or blue pigments called biliproteins, which are divided into classes based on the position of their absorption peaks. The phytoplankton absorption coefficient describes the spectral absorption for natural waters comprised of mixtures of phytoplankton and has been commonly parameterized by chlorophyll concentration and dominant cell size [15, 16].

Scattering processes, which include *refraction*, *reflection* and *diffraction*, occur at the boundary of a particle with a different *index of refraction*, the ratio of the speed of light in the surrounding medium to the speed of light within the particle, than the surrounding medium. Scattering is predominantly elastic, the energy of the photon is conserved, but the direction of propagation is altered. Rather than reducing light, scattering works to inhibit the straight-path vertical penetration of light. The total scattering coefficient (b) can be subdivided into light which scatters in the forward direction (b_f) (0–90°) and the backward direction (b_b) (90–180°) relative to the unattenuated beam. The backscattered light is the radiance that is scattered out of the water column and measured by a sensor as "ocean color." The magnitude of b_b is a function of the concentration, composition (i.e., index of refraction), shape, and size of particles [17].

Water molecules, salts, organic and inorganic particles, and bubbles provide strong contributions to light scattering in the ocean. Scattering by pure water is the result of density fluctuations from the random motion of water molecules and has a wavelength dependence of λ^{-4} [18]. The presence of salt increases scattering, where pure seawater, with a salinity of 35–38‰, scatters 30% more light than pure water devoid of salt. When particles are present, as in natural waters, scattering increases markedly [19]. The scattering coefficient for the clearest surface waters is an order of magnitude greater than that of pure seawater. Particles that are large relative to the wavelength of light scatter mainly in the forward direction via diffraction, where photons propagating along the particle boundary change their direction in response to the boundary in a manner proportional to the cross-sectional area of the particle. Photons entering large particles are likely absorbed. Conversely, small particles mainly reflect and refract light in a manner proportional to the volume of the particle. Small particles with an index of refraction that deviates markedly from 1, including micron (10^{-6} m)-sized calcium carbonate plates or coccoliths generated by coccolithophorid phytoplankton ($n' = 1.25$) or bubbles ($n' = 0.75$), are highly efficient at scattering light in the backward direction [17].

The processes of absorption and scattering are considered additive, therefore the sum of the contribution of each constituent determines the magnitude of the total coefficients a_t and b_t. As such, IOPs are separated into operationally defined components which comprise a and b_b:

$$a_t = a_w + a_{ph} + a_d + a_g, \text{ and}$$
$$b_{bt} = b_{bw} + b_{bp}$$

where the subscripts correspond to water (w), algal or phytoplanktonic (ph), non-algal or detrital (d) matter, and dissolved material, originally termed "gelbstoff" (g). Dissolved material does not scatter light and the contributions of both algal and non-algal matter are generally consolidated into backscattering from particulate (p) material. Recent advances in optical instrumentation have allowed for the measurement of absorption and scattering properties in situ and contributed to advances in ocean color remote sensing [20].

Apparent Optical Properties

Measurements of how light of different wavelengths attenuates with depth in the water column have been the historical basis of optical oceanography [21] following from the use of white Secchi disks to estimate water clarity. The properties that can be derived from measurements of ambient light in the water column are generally termed "apparent" optical properties (AOP) because they operate as optical properties describing the fundamental properties of the medium with only a slight dependence on the angular distribution of the light field. Spectral radiance, L, is the fundamental radiometric quantity which describes the spatial, temporal, directional, and wavelength-dependent structure of the light field in units of radiant flux per area per wavelength per solid angle (W m^{-2} nm^{-1} sr^{-1}) [18]. Planar downwelling irradiance, E_d, is a measure of the radiant energy flux incident on the surface from all directions or solid angles contained in the upper hemisphere, with units of radiant flux per unit area per unit wavelength (W m^{-2} nm^{-1}). The same concept, applied to the lower hemisphere, describes upwelling irradiance, E_u. The ratio of the upwelling to downwelling irradiance yields *irradiance reflectance*, R, a measure of how much light of a certain wavelength entering the ocean is scattered backward by ocean molecules and particles.

For remote sensing purposes, only the radiance from a specific direction is measured by a sensor, not the entire upwelling irradiance. Hence, the color is parameterized as *remote sensing reflectance* (R_{rs}, sr^{-1}), which is the ratio of water-leaving radiance to downwelling irradiance. The term "water-leaving radiance" represents the radiance signal emerging from the water column in a nadir direction and specifically excludes those upward-directed photons that have only reflected off the sea surface and not penetrated the water column (i.e., sun glint). The term R_{rs} represents the proportion of the downwelling light incident on the water surface that is returned through the air-water interface in the nadir direction due to differential absorption and scattering processes. The parameter R_{rs} is proportional to backscattering coefficient and inversely proportional to absorption coefficient and can be approximated as:

$$R_{rs} = \frac{f}{Q}\frac{b_b}{(a + b_b)}$$

where the ratio f/Q is related to the bidirectionality of the light field and varies from 0.09 to 0.11 for most remote sensing applications [22].

The rate of change of radiance and irradiance with depth, known as the vertical diffuse attenuation coefficient (K; m^{-1}), is another principle AOP. Irradiance and radiance decrease approximately exponentially with depth. The downward diffuse attenuation coefficient, K_d, the rate of decrease in downwelling irradiance, $E_d(0)$, with depth (z),

$$E_d(z) = E_d(0)e^{-K_d z}$$

is commonly used in biological studies and is closely linked to the absorption coefficient of the medium specifically. The optical depth, ζ, corresponding to any given physical depth is defined below:

$$\zeta = K_d z$$

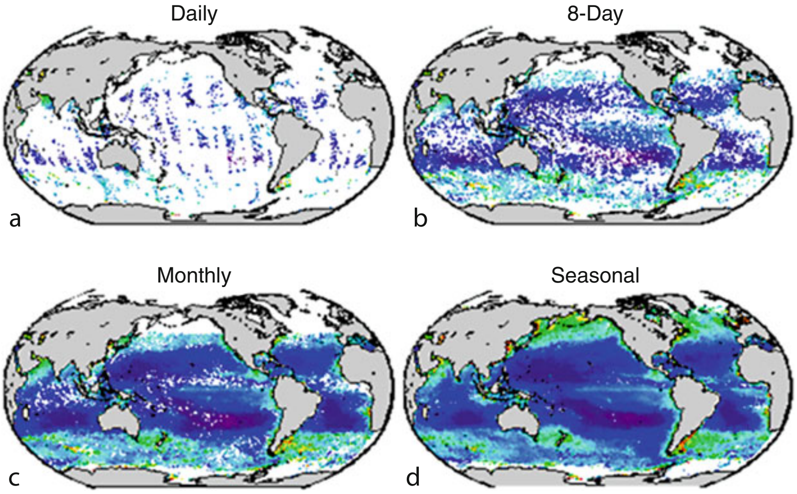
Optical depths frequently used by biologists include 2.3 and 4.6, corresponding to the 10% and 1% light levels, respectively. Also, the portion of the surface water column contributing 90% of the water-leaving radiance has a depth, z, described by $z = 1/K_d$ [12]. The radiative transfer equation is the mathematical formulation that defines the relationship between the apparent and inherent optical properties of natural water bodies [18] and is the basis for the semi-analytical models used in ocean remote sensing.

Basics of Ocean Color Remote Sensing

Many challenges are inherent to remote sensing of ocean color. In comparison to land, the ocean target is dark, with an albedo of only a few percent. This means that most of the light that enters the water is propagated downward into the water column and only a few percent is scattered back out again. This is quite different from land and ice surfaces which have a much higher albedo. Most ocean color sensors are passive in that they measure only the radiation that originates from the sun, as opposed to active sensors that produce and sense their own stream of light (e.g., Light Detection and Ranging or LIDAR). Viewed from space, moreover, the ocean is observed through a thick atmosphere which reflects sunlight back to the sensor and is significantly brighter in the visible wavelengths than the water itself. In technical terms, this is quantified as a low signal-to-noise ratio where the "signal" is the light reflected from within the ocean and the "noise" is light reflected from the atmosphere and sea surface. This section outlines the platforms, calibration, atmospheric correction, and levels of data processing critical for successful ocean color remote sensing.

Sensors and Platforms

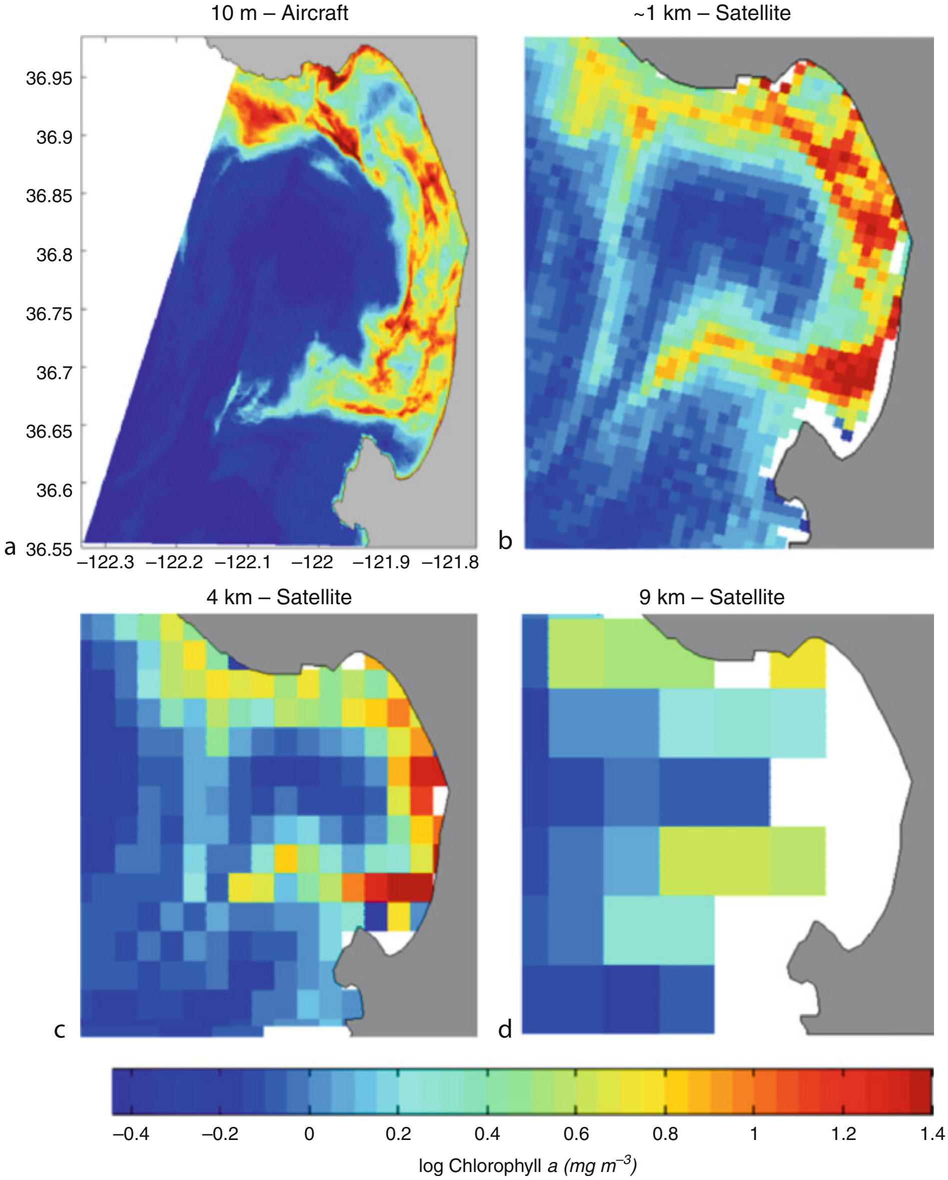
Ocean color sensors can be mounted on space-based satellites or on suborbital platforms like aircraft or unmanned aerial vehicles. The spatial and temporal sampling and the questions that can be addressed with the data depend on the type of platform employed. Most current ocean color sensors have a wide field of view, which translates to a wide sampling swath, and are mounted on sun synchronous polar-orbiting satellites (e.g., CZCS, SeaWiFS, MODIS Aqua and Terra). These sensors have the potential to provide global coverage of the earth roughly every 3 days at the equator and more frequently at the poles. However, clouds obscure the ability of the sensor to view the ocean color and, in reality, temporal sampling for any given region is much less. Data are frequently averaged over longer time periods to produce weekly, monthly, and seasonal composite images of the global ocean (Fig. 2). The spatial resolution is also limited nominally to 1 km pixel widths (and down to 250 m for select channels) in these polar-orbiting sensors in part because of limitations in signal-to-noise inherent to the dark ocean surfaces (see atmosphere correction below). Global datasets are often aggregated to 4-km



Remote Sensing of Ocean Color. Figure 2
Global maps of satellite-derived chlorophyll showing increasing levels of temporal resolution from daily to seasonal. Imagery from MODIS Aqua satellite from 2006: (**a**) 17 December; (**b**) 11–17 December; (**c**) 1–31 December; (**d**) Autumn. White spacing in imagery represents gaps in orbital coverage (daily image), as well as clouds and ice cover. Merging of imagery from different sensors can provide enhanced daily coverage [100]

or 9-km pixels. However, higher spatial resolution on the scale of meters can be obtained from some space-based platforms and from ocean color sensors placed on aircraft (Fig. 3).

The current suite of ocean color sensors has nominally six to seven spectral bands spanning the visible wavelengths (400–700 nm). These bands are not spread uniformly across the visible spectrum, but have been selected to correspond to reflectance characteristics of open ocean waters, particularly those related to phytoplankton pigment absorption features. Three bands are generally found in the "blue" (near 410, 440, and 490 nm), one to two bands in the "green" (510 or 530, 560 nm), and one to two channels in the



Remote Sensing of Ocean Color. Figure 3
Ocean color remote sensing imagery of Monterey Bay, California, illustrates different spatial resolutions available: (**a**) AVIRIS sensor flown on an aircraft, 10 m pixels [25]; (**b**) SeaWiFS satellite Level 2 data, 1 km pixels; (**c**) SeaWiFS satellite gridded to 4-km pixels; (**d**) SeaWiFS satellite Level 3 9-km standard product

"red" (670, 680 nm). In addition, channels are also incorporated in the near infrared (NIR) to short-wave infrared (SWIR) for purposes of atmospheric correction (see section "Atmospheric Correction"). Most of the visible channels were selected to match absorption features of phytoplankton and other constituents. Additional channels are also needed to bridge the large 100 nm gap between 560 and 670 nm, where absorption features are dominated by water, to better constrain backscattering in complex coastal waters [23, 24]. New technology has allowed for the development of sensors that span the full range of visible and near infrared (NIR) spectrum or "hyperpsectral," also referred to as imaging spectrometers.

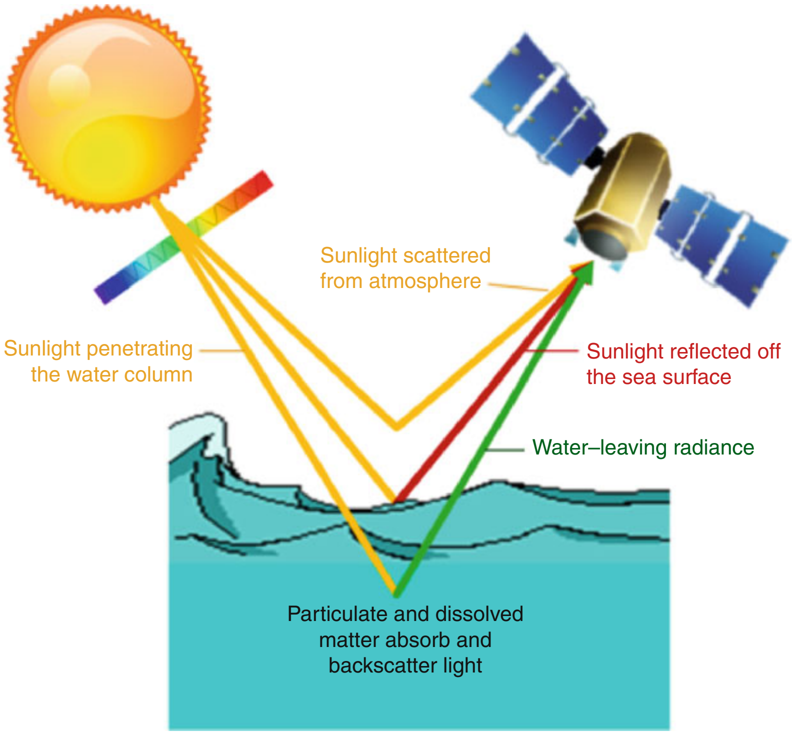
No single platform is ideal for addressing all of the temporal and spatial variability in the oceans. A constellation of ocean color imagers with complementary capabilities and specifications is ultimately required to adequately address the diverse requirements of the coastal research and applied user communities. For example, the Hyperspectral Imager for the Coastal Ocean (HICO) was recently installed on the International Space Station for the study of the coastal ocean and adjacent lands. This imaging spectrometer is intended to provide hyperspectral imagery at 100-m resolution sampling at different angles and times of the day for selected regions. Sensors are also being considered for placement on geostationary satellites, similar to the international constellation of meteorological satellites. Such sensors would look at the same regional location on earth for extended periods of time and be able to provide better temporal resolution of ocean processes and episodic hazards. Regional efforts such as the Geostationary Ocean Color Imager (GOCI) on the COMS-1 platform from South Korea are already planned for launch. In addition, higher spatial and spectral resolution polar orbiting sensors are proposed to address questions related to seasonal variability in global coastal habitats and polar ice cover.

Portable sensors flown on aircraft or unmanned aerial vehicles (UAV's) provide a critical sampling niche distinct from satellite-borne sensors that is particularly well suited for coastal applications and ice research (Fig. 3a) [25]. Airborne sensors can sample at finer spatial scales (meters), can operate under clouds and with nearly unlimited repeat coverage, and are effective platforms for high-resolution active sensors (e.g., LIDAR). Flight lines and scanning geometries can also be oriented to avoid sun glint and their range can be greatly expanded by launching from ships. The technology required to build portable sensors for coastal applications is developing with wide field of views, minimum polarization dependence, high response uniformity, and optimized signal-to-noise ratio for low-light channels [26, 27]. These sensors are becoming more popular for use in the environmental management of coral reefs, seagrasses, kelps, and other coastal targets, and have the potential to monitor episodic events such as harmful algal blooms and runoff and flooding from storms.

Ocean color sensors in space have traditionally been "whisk broom" in design where a single detector collects data one pixel at a time as the telescope rotates to build up pixels along a scan line. Some satellites and most of the suborbital sensors are "pushbroom" where the entire scan line is imaged synoptically by a line of sensors arranged perpendicularly to the flight direction. In order to achieve high-quality data that can track climatological trends in ocean color, sensors are required to have very high radiometric accuracy and stability. Detectors are calibrated pre- and post-launch and degradation over time is carefully quantified with vicarious calibrations from field measurements and ideally lunar imaging. Periodic reprocessing of the satellite data is considered critical to obtaining high-quality datasets and continuity over multiple missions [5, 28].

Atmospheric Correction

One of the most challenging aspects of ocean color remote sensing is successfully removing the atmospheric signal from the water column signal. Aerosols and gas molecules are the primary contributors to the radiance measured at the top of the atmosphere. Approximately 80–85% of the radiance measured at the sensor is the result of Rayleigh scattering by molecules in the atmosphere that are small relative to the wavelength of light. Photons reaching the sensor (L_u) are a combination of those scattered by the atmosphere (L_p), reflected at the air-water interface (L_r), known as specular reflection, or have been backscattered from within the water column, dubbed water leaving radiance, or L_w (Fig. 4). The water-leaving radiance, used for most ocean color applications, is only a small portion of the signal retrieved at a satellite and must be


Sunlight scattered from atmosphere
Sunlight penetrating the water column
Sunlight reflected off the sea surface
Water–leaving radiance
Particulate and dissolved matter absorb and backscatter light

Remote Sensing of Ocean Color. Figure 4
Radiance measured by a satellite includes light scattered by the atmosphere and reflected off the sea surface (i.e., glint). In a process called "atmospheric correction," these signals are removed leaving the "water-leaving radiance" or the light that has penetrated the water column and been backscattered out to the satellite – a measure of ocean color

differentiated from the photons scattered within the atmosphere and specularly from the sea surface in a process called "atmospheric correction."

Rayleigh scattering, which decreases with wavelength (λ) following λ^{-4}, can be estimated using a single-scattering radiative transfer equation using the atmospheric pressure and appropriate viewing geometry [29]. An additional 0–10% of the radiance signal is due to aerosols (i.e., haze, dust, and pollution), particles with sizes comparable to the wavelength of light which absorb and scatter as a complex function of their type, size, and concentration. The type and concentrations of aerosols overlying the ocean are quite variable in space and time, particularly in coastal regions subject to urban pollution and terrestrial dust [30].

Atmospheric correction of aerosols remains a challenge for accurately deriving water-leaving radiance from satellites and aircraft. Approaches generally focus on channels in the NIR and even in the short wave infrared (SWIR) [29, 31, 32]. Because water absorbs so heavily in the infrared, very few photons are reflected out of water in this part of the electromagnetic spectrum and the signal is dominated by reflection from atmospheric gases and aerosols. Various types of models are used, including coupled models and multi-scattering models, to infer the contribution of aerosol reflectance in the visible portion of the spectrum from the infrared. Aerosol reflectance is not spectrally flat, but varies with wavelength, and at least two channels are necessary to determine the spectral shape of aerosol reflectance and extrapolate from the NIR to visible wavelengths [29, 33].

Dust, particularly from desert storms, can also impact the optical properties of the atmosphere and most atmospheric correction algorithms for ocean color sensors are not capable of handling absorbing mineral dust (i.e., colored dust) [34]. For example, airborne plumes of Saharan dust are observable all year on satellite images over the Tropical Atlantic and may be increasing in areas like the Mediterranean Sea [35]. If colored dusts are not properly corrected for in the atmospheric correction schemes, then the color of the ocean is not accurately estimated resulting in errors in chlorophyll and other biogeochemical properties retrieved from the satellite data [36]. In addition to its radiative impact, it has been suggested that this mineral dust has a substantial influence on the marine productivity and may also carry pollutants to the oceans [37, 38].

Whitecaps breaking on the sea surface must also be corrected from derivations of water-leaving radiance. Whitecap reflectance is often modeled using an empirical cubic relationship to wind speed and an approximate reflectance value for an individual whitecap [39], but such models often overcorrected the imagery, and a fixed whitecap correction is applied when wind speeds exceed a threshold (e.g., 8 m s^{-1} for SeaWiFS). At high winds, some of the signal attributable to whitecaps is removed by the aerosol corrections.

Levels of Processing

Standards for ocean color data processing, developed at US National Aeronautics and Space Administration (NASA) for the SeaWiFS mission [40], are widely followed by the international community of ocean color users and involve four levels of processing (Table 1).

Ocean Color Algorithms

This section presents the classification of the global ocean into two optical classes: Case 1 and Case 2. The general approaches for two of the main products from ocean color imagery, chlorophyll and primary productivity, for Case 1 waters and a description of the semi-analytical algorithms used for both Case 1 and Case 2 waters are presented.

Optical Classification of Aquatic Systems

Ocean waters have long been classified based on their color properties [41]. A classification system introduced in 1977 differentiates phytoplankton-dominated waters from those where inorganic particles are dominant, known as Case 1 and Case 2, respectively [42].

Remote Sensing of Ocean Color. Table 1 Levels of data processing products from ocean color satellites

Level	Processing	Spatial qualities
0	Raw data as measured directly from the spacecraft	Satellite coordinates at highest spatial resolution
1	Converted to radiance using calibrations and sensor characterization information	Satellite coordinates at highest spatial resolution
2	Atmospherically corrected to water-leaving radiance and derived products	Satellite coordinates at highest spatial resolution
3	Derived products have been mapped onto a two-dimensional grid at known spatial resolution and can be averaged over timescales (weekly, monthly)	Regular gridded data at lower spatial resolution (e.g., 4 or 9 km)
4	Products that have been merged or assimilated with data from other sensors, in situ observations, or model outputs	Regular gridded data at lower spatial resolution

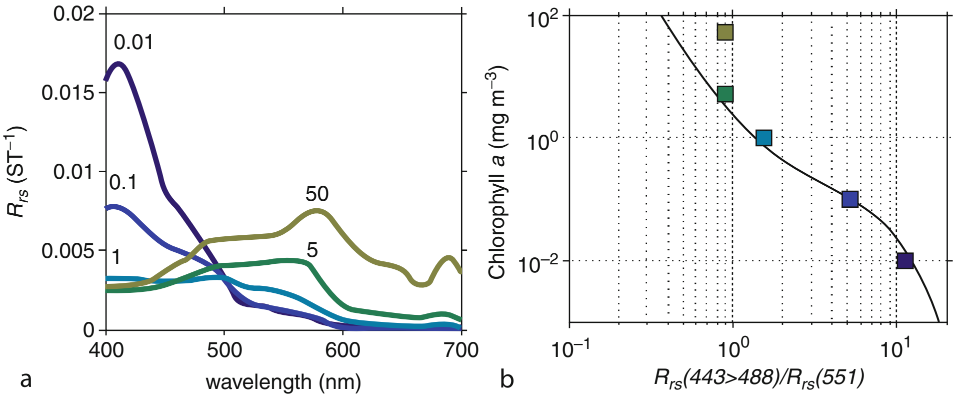
These cases have evolved from their original forms into the categories used today: Case 1 waters are those waters where optical properties are determined primarily by phytoplankton and related colored dissolved organic matter (CDOM) and detritus degradation products; Case 2 waters are waters where optical properties are significantly influenced by other constituents such as mineral particles, CDOM, or microbubbles that do not covary with the phytoplankton concentration [8, 43]. In today's world, approximately 97% of the surface ocean falls toward the optically simple, deep water, Case 1 classification. When inorganic, organic, particulate, and dissolved material all vary independently of one another, such as in coastal ecosystems with considerable riverine influence, bottom resuspension, or optically shallow regions, the system falls toward the Case 2 classification, also called "optically complex."

This binary classification scheme has been prevalent in bio-optical modeling of ocean waters and development of ocean color algorithms. However, many problems exist with use of such simplified schemes in modeling natural systems. For example, there is no sharp dividing line between the cases and each investigation tends to use as different criteria for defining Case 1 and Case 2. Commonly the two cases are defined by the relationship between chlorophyll and remote sensing reflectance or scattering. Even in the global ocean considered to be Case 1, CDOM concentrations do not covary with the instantaneous chlorophyll concentration [44], but can vary from 30% to 60% of the total non-water light absorption [45] and result from differences in water mass ventilation, water column oxidative remineralization, and photobleaching [46].

In *optically shallow waters*, in addition to the water column and its constituents (i.e., dissolved and particulate material), the bottom contributes to the water leaving radiance in a way that depends on the bottom composition and roughness. Periodic measurements of bottom types using passive remote sensing in coastal systems are valuable for describing and monitoring habitats [47]. The magnitude and spectral quality of light reflected off of the bottom material can allow separation of bottom reflectance from the water column signal, where different bottom types will have a different effect on reflectance. Shallow, clear water will yield the most information about bottom material, more readily allowing spectral discrimination of bottom type. However, as depth and the diffuse attenuation coefficient, K_d, increase, the bottom signal becomes difficult to differentiate.

Empirical Chlorophyll Algorithms

Standard calculation of chlorophyll from ocean color imagery involves an empirical relationship developed from field observations collected throughout the global ocean [10]. Algorithms are typically not developed from the remotely sensing imagery itself, because this would incorporate any biases in calibration and atmospheric correction procedures used to derive reflectance, as well as any spatial inhomogeneity in parameters over pixel scales, and would require new algorithms for every new calibration, reprocessing, and sensor. Empirical solutions are used because an analytical solution to the problem requires an assessment of the entire radiance distribution and depth derivative and such measurements are not possible with remote sensing [48]. Only the upward flux incident upon the water-air interface at angles less than 48°, the angle at



Remote Sensing of Ocean Color. Figure 5
(**a**) Remote sensing reflectance (R_{rs}) spectra modeled for different concentrations of chlorophyll *a* (Chl) from 0.01 to 50 mg m^{-3}. The color of each line represents the modeled ocean color a human might observe following [61]. (**b**) The empirical OC3M model for deriving Chl from R_{rs} for the MODIS Aqua sensor. The model uses the "blue" channel with the highest R_{rs} value (443 or 488 nm) divided by the "green" channel at 551 nm. Each *square* represents the modeled Chl for the corresponding R_{rs} spectra in panel A and demonstrates how the model becomes less accurate at high Chl

which complete internal reflection occurs, is measurable from above the sea surface [6] and generally only the flux emitted in a single viewing angle is remotely sensed.

The current empirical algorithms use the shift in ocean color from "blue" at low Chl, where R_{rs} peaks at 400 nm, to "green" at high chlorophyll, where R_{rs} peaks at 555 nm (Fig. 5a). Empirical ocean color algorithms have been applied to the vast majority of the global ocean considered Case 1 and use multiple ocean color bands typically log-transformed and in a ratio formulation to minimize problems with atmospheric correction and differential scattering in the ocean. The coefficients for the algorithms are regularly adjusted to account for different sets of wavebands in various sensors and as new field data becomes available (Table 2). The OC3M algorithm developed for MODIS, for example, uses a 4th order polynomial derived from a large global dataset of field measurements of chlorophyll and R_{rs}. It uses a logarithmic ratio of blue light (either 443 and 488 nm depending on which is greater) to green light (555 nm) and follows an inverse relationship such that low Chl is retrieved or high ratios when the ocean color is blue and high Chl when more green light is reflected (Fig. 5b). These types of algorithms tend to work best at lower Chl (<1 mg m^{-3}), found in most of the world ocean, where the algorithm has a flatter slope [49].

For much of the open ocean where chlorophyll concentrations are low, the empirical algorithms work well and relative error is estimated to under 35% [50]. However, empirical derivations of chlorophyll in Case 1 waters can be in error by a factor of 5 or more, particularly at higher Chl [49]. Such variability is due to differences in absorption and backscattering properties of phytoplankton and related concentrations of colored dissolved organic matter (CDOM) and minerals. The empirical algorithms have built-in assumptions that follow the basic precept of biological oceanography; i.e., oligotrophic regions with low phytoplankton biomass are populated with small phytoplankton while more productive regions contain larger bloom-forming phytoplankton. With a changing world ocean, phytoplankton composition may shift in response to altered environmental forcing and CDOM and mineral concentrations may become uncoupled from phytoplankton stocks creating further uncertainty and error in the empirical approaches [49].

The empirical approach is not widely applicable in Case 2 waters, generally found near the coasts. Such waters are influenced by freshwater plumes with CDOM and minerals that significantly impact the

Remote Sensing of Ocean Color. Table 2 Empirical chlorophyll algorithms for a variety of ocean color sensors

Name[a]	Sensor	Channels[b]		Coefficients[c]				
		Blue	Green	a0[c]	a1	a2	a3	a4
OC4	SeaWiFS	443 > 490 > 510	555	0.366	−3.067	1.93	0.649	−1.532
OC3S	SeaWiFS	443 > 490	555	0.2409	−2.4768	1.5296	0.1061	−1.1077
OC2S	SeaWiFS	490	555	0.2372	−2.4541	1.7114	−0.3399	−2.788
OC3M	MODIS	443 > 488	551	0.283	−2.753	1.457	0.659	−1.403
OC2M	HMODIS	469	555	0.1543	−1.9764	1.0704	−0.2327	−1.1404
OC4O	OCTS	443 > 490 > 520	565	0.4006	−3.1247	3.1041	−1.4179	−0.3654
OC3O	OCTS	443 > 490	565	0.2836	−2.1982	1.0541	0.186	−0.717
OC2O	OCTS	490	565	0.2805	−2.167	1.1789	−0.1597	−1.5591
OC3C	CZCS	443 > 520	550	0.3012	−4.4988	9.0983	−9.9821	3.235

[a] Name of ocean color (OC) algorithm incorporates the number of wavebands (2–4) used in the formulation and the initial for the sensor used (S = SeaWiFS; M = MODIS; O = OCTS; C = CZCS)

[b] The algorithms use a log-transformed ratio of "Blue" (443–520 nm) to "Green" (550–565 nm) remote sensing reflectance (R_{rs}). When more than one "Blue" channel is provided, only the channel with the highest R_{rs} is used. x = log10(R_{rs}(Blue)/R_{rs}(Green))

[c] Chlorophyll *a* is modeled as a fourth polynomial fit to the field data such that: Chl $= 10^\wedge(a0 + a1{}^*x + a2{}^*x^2 + a3{}^*x^3 + a4{}^*x^4)$

optical properties, as well as resuspension of bottom sediments [51]. Phytoplankton assemblages can also be diverse in coastal regimes and light absorption per unit of Chl is difficult to constrain. Melting and runoff of glacial sources can increase particle concentrations in the nearshore and change phytoplankton assemblages. In order to use remote sensing in coastal waters, semi-analytical models are employed that are able to decompose the reflected color into the many absorbing and scattering constituents in the water column (see section "Semi-analytical Algorithms").

Primary Productivity Algorithms

Net primary production is a key parameter derived from ocean color data that provides a measure of how much carbon dioxide is taken up and incorporated into ocean phytoplankton during photosynthesis. Export of fixed carbon to the ocean interior, while only a fraction of the total biomass produced, provides a long-term sink for atmospheric carbon dioxide [52]. While satellite-derived Chl is not a direct measure of carbon fixation in phytoplankton, such estimates are typically derived from correlates of Chl and rates of carbon fixation [53]. Net primary productivity varies with phytoplankton species assemblages and their physiological state related to light, temperature, nutrients, and other environmental factors.

A variety of formulations have been developed for ocean color remote sensing and parameterized for the global ocean or specific regions. Models are generally restricted to parameters that can also be globally derived from remote sensing imagery, such as sea surface temperature and photosynthetically available radiation (PAR). Moving from a standing stock of phytoplankton biomass to photosynthetic rate requires a time-dependent variable. Solar radiation in the form of PAR is commonly used in formulations to convert biomass to primary productivity. The physiological response of the measured chlorophyll to light, nutrients, temperature, and other environmental variables must also be incorporated in the model. Primary productivity models can be differentiated by the degree of explicit resolution in depth and irradiance [53].

Round robin experiments have been conducted to compare the performance of models for assessing global productivity from ocean color imagery, as well as the output from ecosystem-based general circulation models [1, 54]. The third such effort found that global average primary productivity varied by a factor of two between models and the global mean productivity for the different model groups ranged from 44 to 57 Gt C year^{-1} with an average of 50.7 Gt C year^{-1}. The models diverged the most in the high-nutrient low chlorophyll waters of the Southern Ocean. Primary productivity algorithms have also been formulated from remote sensing estimates of the inherent optical properties (such as light absorption and backscattering) directly [55, 56], without incorporating Chl and the associated uncertainties inherent in that parameter.

Semi-analytical Algorithms

The empirical algorithms used for deriving chlorophyll have been likened to a "black box" that provides no mechanistic understanding of ocean optics and are particularly challenging to apply in a changing ocean, when the water properties are different from the empirical data used to develop the formulation [57]. Analytical solutions to deriving IOPs from water-leaving radiance are not possible because the radiance can only be measured from a few angles. Semi-analytical algorithms (or "quasi-analytical") are based on a fundamental understanding of the propagation of light in the ocean and provide a more mechanistic approach to ocean color. These algorithms incorporate some empirical approximations, but do not rely on fixed predetermined relationships between the absorption and backscattering components of the water column.

In semi-analytic models, the ocean color signal is inverted to obtain estimates of the various absorbing and backscattering constituents directly. Parameterization of how water, phytoplankton, and dissolved and detrital material inherently absorb and backscatter light across the visible spectrum (i.e., their spectral shape) is used in these models. The spectral reflectance measured at the satellite is often inverted to retrieve the amounts of each individual component contributing to the absorption and backscattering of light. Such algorithms are the primary methods for obtaining CDOM distributions across the ocean surface [58]. In semi-analytical models, the biogeochemical parameters, such as Chl and total suspended matter, are derived secondarily from the IOPs. Semi-analytical formulations vary in terms of their architecture and statistical methods employed to retrieve the inherent optical properties from the remote sensing signal and the empirical parameterizations within the models [57].

Applications for Oceanography

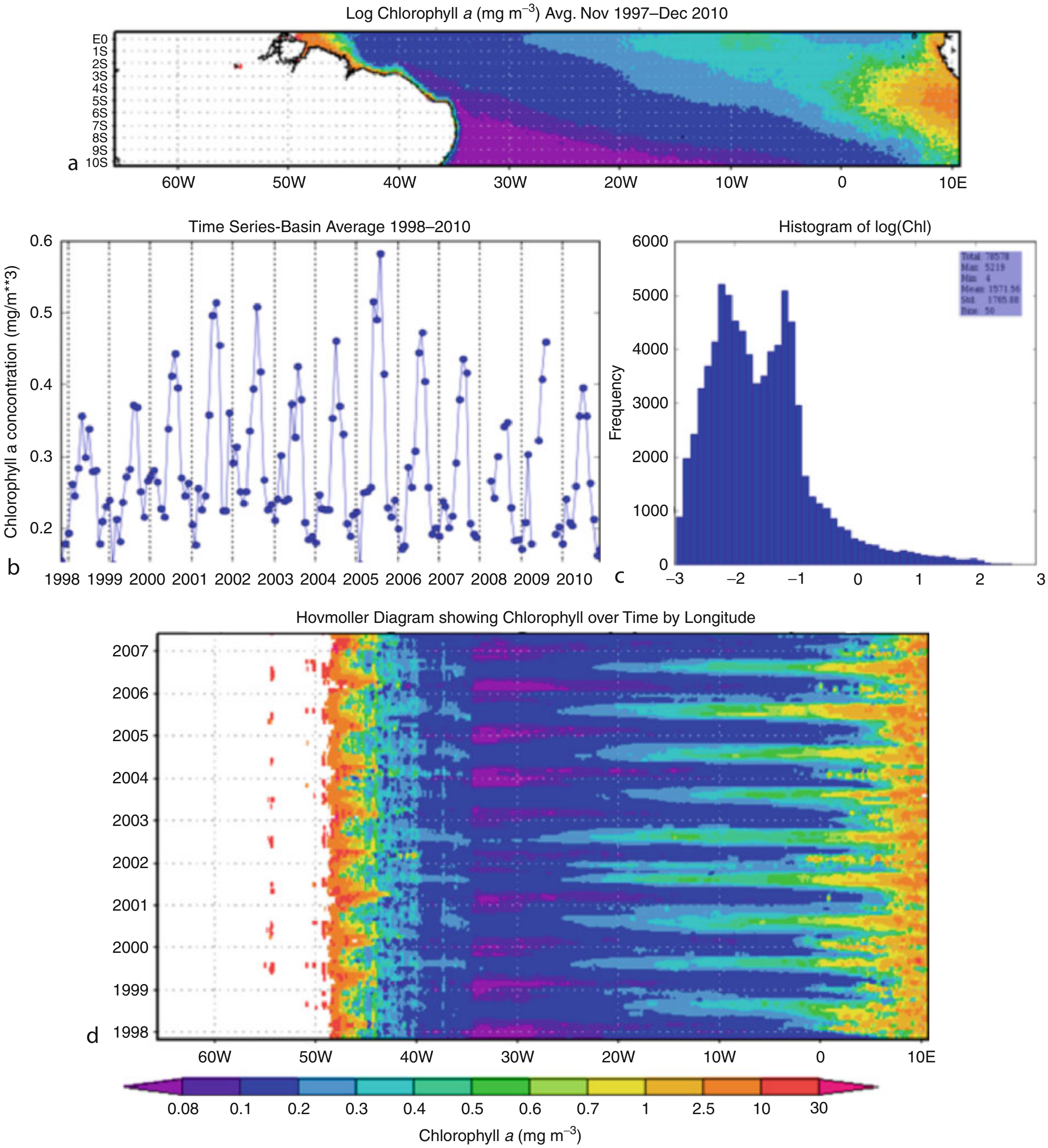
Ocean color remote sensing is an important tool for many branches of oceanography, including biological, physical, and chemical oceanography. The section below summarizes only some of the main applications of ocean color remote sensing with the understanding that the uses of ocean color are continuously expanding. A recent monograph from the International Ocean Color Coordinating Group (IOCCG) entitled "Why Ocean Colour?: The Societal Benefits of Ocean-Colour Technology" extensively documents the many uses of ocean color remote sensing from scientists to environmental managers to the general public [7]. Web-based software has also been developed, see, e.g., Giovanni [59], which allows the public to freely map and analyze ocean color imagery over time and space. Figure 6 provides an example of various types of figures that can be easily generated from remotely sensed chlorophyll using that software.

Biological Oceanography

Apart from estimating chlorophyll and primary productivity, ocean color remote sensing has many biological applications that range from phytoplankton physiology to assessing distributions of migrating whales. Phytoplankton physiology, particularly the efficiency of light capture and utilization, has been modeled from the natural fluorescence signature provided by ocean color remote sensing [60]. Even though the spectral resolution available in most current ocean color satellite is limited to six to eight available spectral channels [61], a variety of phytoplankton taxa and groups have also been distinguished from satellite imagery based on their unique optical properties and/or regional tuning of algorithms using knowledge of the local phytoplankton composition. Phytoplankton taxa can have unique sets of accessory pigments that differentiate them from one another and can result in unique absorbance spectra. In addition, phytoplankton can have cell walls or exterior plates comprised of different materials (e.g., silica, calcium carbonate) that can make them more or less reflective. Various approaches have been developed to map size classes (from pico- to microplankton) or major groups of phytoplankton in the global ocean [62]. Other algorithms have targeted particular phytoplankton taxa such as coccolithophores, nitrogen-fixing *Trichodesmium* [63], toxic dinoflagellates [64], and nuisance cyanobacteria [65].

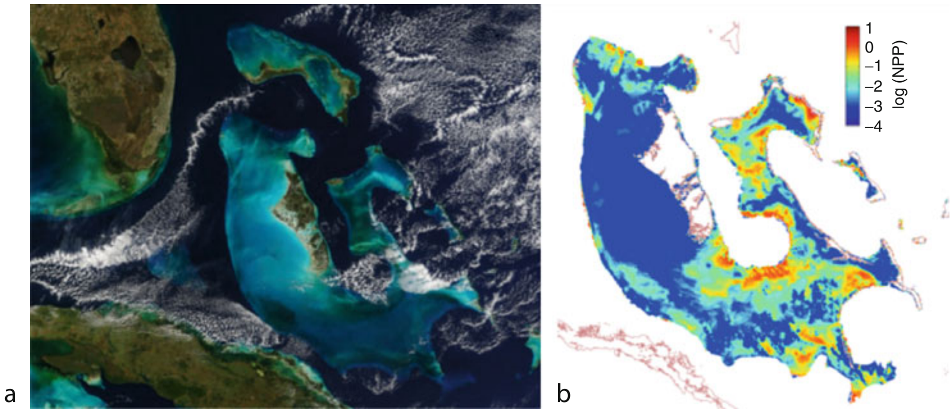
Satellite-derived chlorophyll and primary productivity provide a key metric to assess marine ecosystems temporally on a global scale and have been used extensively to monitor conditions that impact other biological organisms in the sea. The relationship between satellite-derived chlorophyll data and organisms at higher trophic levels depends upon the number of linkages in the food web. For species like anchovies and sardines, which eat phytoplankton in their life cycle, the linkage can be direct [66]; whereas, many trophic levels can exist for other species and the relationship can be quite nonlinear [7]. The distribution, movement, and migration of whales, dolphins, pinnipeds, penguins, and sea turtles has been related, either directly or indirectly, to remotely sensed patterns of Chl (reviewed in [7]). Most fish have planktonic larval stages that are strongly influenced by ocean circulation and recruitment success has been found to be related to the degree of timing between spawning and the seasonal phytoplankton bloom, as observed from satellites [67]. Ocean color remote sensing has also been used to study invertebrates in the global ocean, such as shrimp in the Newfoundland-Labrador Shelf [68] and pteropods and pelagic mollusks in the Ross Sea [69]. Mean net primary productivity, determined from ocean color satellite imagery, elucidates species richness in biogeographical studies of cephalopods [70].

New techniques have also been developed to use ocean color remote sensing in optically shallow water systems to deduce changes in benthic habitats [71]. Optically shallow water occurs when the seafloor contributes to the reflectance signal observed remotely by a satellite (Fig. 7a) and is defined by a combination of water clarity, water depth, and bottom composition.



Remote Sensing of Ocean Color. Figure 6

Various times series analyses that can be conducted with standard Level 3 chlorophyll imagery including (**a**) Temporally averaged spatial distributions; (**b**) time series of interannual variability; (**c**) histograms showing the statistical distributions; (**d**) Hovmoller plots presenting both spatial (x-axis) and temporal (y-axis) variability. Such plots can be easily generated by the public with the Giovanni interface [59]



Remote Sensing of Ocean Color. Figure 7
The Great Bahama Bank is an example of optically shallow water where the seafloor color can be observed from space. (**a**) Pseudo-true color image from MODIS Aqua showing the bright Bahamas Banks with Florida, USA, to the West and Cuba to the Southwest. White wispy clouds can obscure the ocean color. (**b**) Net primary productivity (mgC m^{-2} d^{-1}) of seagrass and benthic algae estimated from ocean color imagery over the Great Bahama Bank [47]

Satellite estimates of biomass and net productivity of seagrasses, kelps, and other benthic producers have been conducted over regional scales [47, 72] (Fig. 7b). Ocean color imagery from aircraft can map fine-scale distributions of seagrasses, coral reefs, and other coastal habitats at local scales [73, 74]. Changes in ocean color signals over time can also be used to assess contributions of coastal carbon to the global carbon cycle [75, 76]. Responses of coastal regions linked to terrestrial changes can also be observed with ocean color imagery. Warming of the Eurasian landmass, for example, has led to enhanced productivity in the water column [77]. Agricultural runoff from fields in Mexico was shown to stimulate large phytoplankton blooms in the Gulf of California that alter water clarity and potentially lead to anoxic conditions [78].

Ocean Physics

Ocean color data is well suited to the detection of convergence zones and oceanic fronts, sometimes better than thermal sensors which penetrate only the skin layer, or the first 10 μm, of the water column. Interestingly, a sequence of ocean-color-derived chlorophyll images may help predict the formation of eddies days before they appear. The increased penetration of visible radiation reveals more frontal features and with greater detail than those retrieved with sea surface temperature data alone [79]. Likewise, upwelling regions, which bring cold, nutrient-rich waters up to the surface can be readily identified in ocean color images as areas with an enhanced chlorophyll concentration. The intensity of upwelling from year-to-year can be tracked through the time series of chlorophyll abundance. Chlorophyll is an effective indicator for detecting anomalous activity in the oceanic environment. Evidence of an El Niño event beginning in November of 1997, during which phytoplankton pigment concentrations appeared anomalously low in the Equatorial Upwelling Zone, was obvious in the continuous coverage supplied by SeaWiFS. The onset of restored upwelling was likewise evident with the increased chlorophyll concentrations during the months of June and July 1998 [80].

Ocean water clarity also affects the distribution of shortwave heating in the water column. Both chlorophyll and CDOM concentrations have been linked to changes in heating of surface waters [81, 82]. Increased clarity would be expected to cool the surface and heat subsurface depths as shortwave radiation penetrates deeper into the water column. Recent studies show that water clarity, as determined from ocean color remote sensing, is an important feature in atmospheric circulation (the Hadley cells), oceanic circulation (Walker Circulation), and formation of mode water [83]. Importantly, ocean color imagery is also critical to predicting tropical cyclone activity. The presence of light-absorbing constituents (like Chl and CDOM) shapes the path of Pacific tropical cyclones and propagation to higher latitudes [84].

Chemical Oceanography

A major contributor to the ocean carbon system is colored dissolved organic material (CDOM), a mixture of compounds produced primarily by decomposition of plant matter. CDOM, when present in high enough concentrations, produces a yellow or brownish color and is highly reactive in the presence of sunlight. When CDOM undergoes photodegradation, organic compounds essential to phytoplankton and bacterial growth are released [85]. Satellite measurements collected using SeaWiFS, MODIS, and MERIS produce daily estimates of CDOM at 1 km resolution. High temporal resolution CDOM maps can be used to identify and track water masses at timescales close to the processes determining its distribution. CDOM dynamics play an important role in ocean biogeochemistry, regulating the absorption of blue and UV radiation in the surface ocean and therefore altering the depth of the euphotic zone [58] and heating surface waters [82]. Although CDOM is difficult to analyze chemically, its distribution and abundance, identifiable using ocean color remote sensing, is highly relevant to understanding carbon cycling in the ocean.

The particulate inorganic carbon (PIC) pool, calcium carbonate ($CaCO_3$), contributes substantially to the ocean carbon cycle and ocean color reflectance. Calcification reduces surface carbonate, decreasing alkalinity. Organic carbon production via photosynthesis counterbalances this effect. Coccolithophores, haptophyte algae, are responsible for the majority of the biogenic particulate inorganic carbon production. Coccolithophores generate and shed tiny white plates of calcium carbonate called coccoliths, which are highly efficient at reflecting light, ultimately producing large turquoise patches in the ocean readily visible in ocean color imagery [86]. Ocean color remote sensing algorithms have been formulated for generating quantitative estimates of particulate inorganic carbon and calcification rates on regional and global scales [87, 88]. A continued, long-term assessment of coccolithophore and particulate inorganic carbon abundance from satellite imagery will aid in understanding the impact of ocean acidification on marine organisms reliant on carbonate for the formation of shells [89].

Ocean color imagery provides the ability to expand small-scale biogeochemical studies to regional or global scales. For example, the marine inorganic carbon cycle has been shown to be not only influenced by marine plankton but also by fish that precipitate carbonates into the surface waters. Extrapolations from satellite-derived net primary productivity up several trophic levels to marine fish [90] reveal that fish may contribute 3–15% of the total oceanic carbon production [91].

Applications for Environmental Monitoring

Ocean color remote sensing plays a major role in monitoring and sustaining the health and resilience of marine ecosystems, including fisheries and endangered species [40]. Ocean color products are helping to address how environmental variability influences annual recruitment of fish stock [92] and to locate and manage fisheries [7]. Ocean color imagery coupled with other remote sensing products such as sea surface temperature is a fundamental tool in ecosystem-based management of marine resources [93].

Ocean color remote sensing can monitor a variety of acute and chronic hazards influencing the oceans including: harmful algal blooms, oil spills, coastal flooding, icebergs and marine debris [7]. A combination of ocean color, field, and meteorological datasets have been critical in identifying the onset of harmful algal blooms (HABs), which can produce toxins and create hypoxic conditions. While toxins cannot be directly observed from ocean color, the onset of potential harmful blooms can be identified using a chlorophyll anomaly method [94] in concert with other forecasting tools such as field and meteorological datasets. This information can then be passed on to coastal managers and state agencies to put strategies in place to deal with an impending bloom. A long-term time series of ocean color products can aid in elucidating forcing and transport mechanisms of these harmful blooms and help improve predictability.

New techniques are being developed for early detection, containment, and clean up of oil spills. Remote sensing can be used to detect oil spills that can change surface reflectance properties and the color of the ocean [95]. Coarse spatial and temporal resolution, limited spectral bands, cloud-cover issues and high sunlight requirements have generally restricted the usefulness of ocean color imagery for oil-spill detection from polar orbiting satellites [96]. Moreover, current processing methods may not allow data availability within hours of data capture. The spatial, temporal, and spectral resolution needed for oil spill recovery planning requires high-resolution, hyperspectral ocean color radiometers deployed in geostationary orbit [40].

Ocean color imagery has also been used to track marine debris on the ocean surface which can entangle a variety of pelagic species, such as endangered sea turtles, seals, and whales. The nets also become ensnared on coral reefs and damage the reef structure and associated organisms that require a healthy reef ecosystem [97, 98]. Satellite ocean color data are part of the methods being developed to locate and identify potential locations of marine debris to aid their removal from these ecosystems.

Ocean color imagery is also useful in monitoring water quality in inland aquatic water bodies. Nuisance algal blooms, such as cyanobacteria, cause aesthetic degradation to lakes and reservoirs resulting in surface scum, unpleasant taste and odor in drinking water (from the production of metabolites such as methyl isoborneol and geosmin), and possible adverse effects to human health from blue-green algal toxins. Predicting the locations and timing of blue-green algal bloom using traditional sampling techniques is difficult and hyperspectral remote sensing can be an important tool in such monitoring efforts [99].

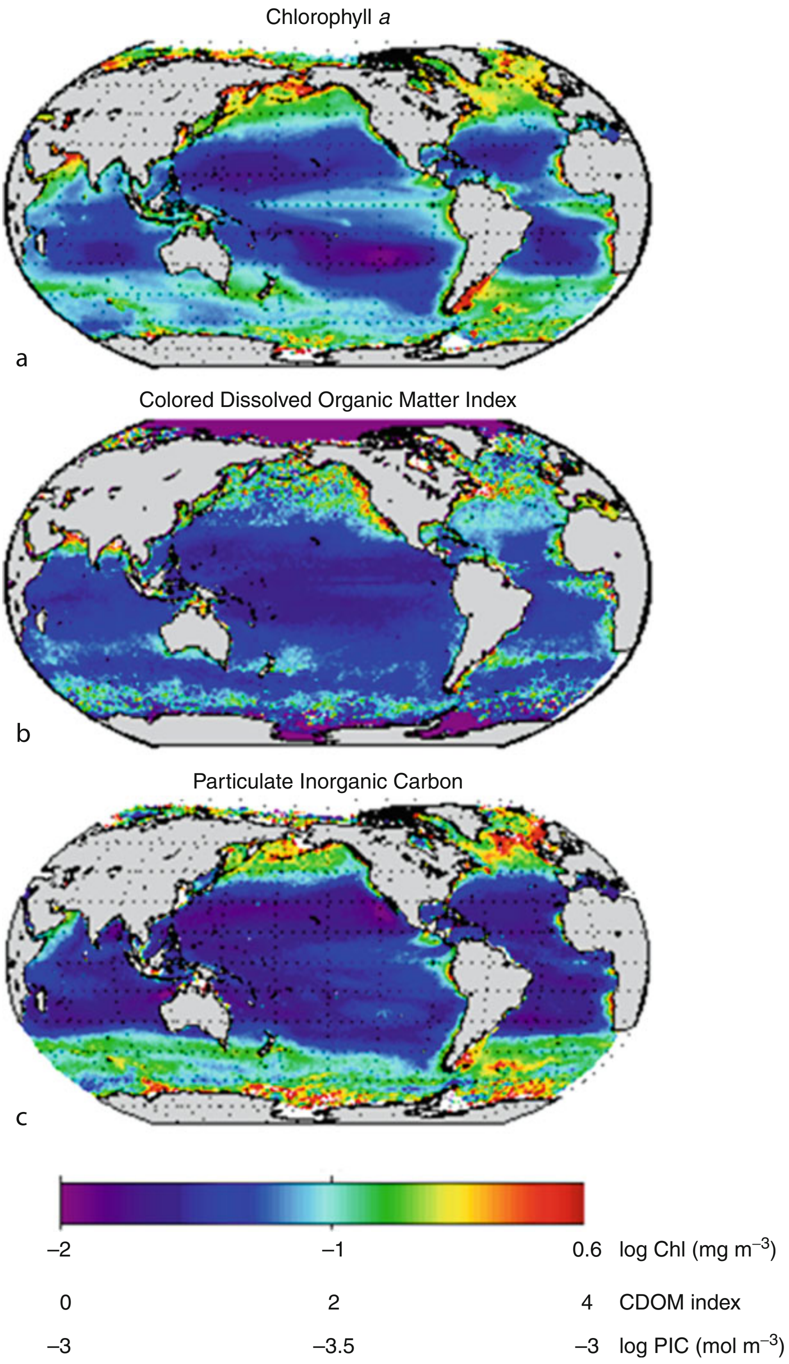
Future Directions

Within a few decades, the ability to view the global ocean color regularly through remote sensing has revolutionized the perceptions about ocean processes and feedbacks to the earth's climate. The decade of continuous ocean color imagery has provided a foundation for assessing change in the earth's systems and long-term averages or "climatologies" of products, such as chlorophyll, CDOM, and PIC, have been produced to provide a baseline of ocean biogeochemistry (Fig. 8). The products obtained from ocean color are now incorporated into all domains of oceanography, global climate forecasts, military applications, and environmental monitoring across the expansive global ocean and the vulnerable coastal regions where most of the human population resides [11]. While successful, the technology and processing of ocean color remote sensing is still in its infancy in terms of monitoring the ocean from immediate to climatological timescales.

The relationships between climatological forcing and biological carbon storage in the ocean are complex and not readily incorporated in models. Ocean color imagery can provide assessments of potential changes to ocean processes including primary productivity, surface heating, sediment plumes, altered food webs, harmful algal blooms, changing acidity, and alterations of benthic habitats in response to shifts in winds and upwelling, clouds and radiative forcing, and storm intensity and frequency. Recent observed changes in chlorophyll, primary production, and the size of the oligotrophic gyres from ocean color satellites are compelling evidence of significant changes in the global ocean. A recent study demonstrates that a time series of at least 40 years in length is needed to unequivocally distinguish a global warming trend from natural variability [6] and sustained long-term observations of ocean color are in jeopardy [40].

In addition to sustained imagery, there is a need for integrating ocean color imagery from different platforms to monitor the oceans and aquatic habitats at a variety of desired spectral, spatial, and temporal resolutions. Integration of satellite sensors with suborbital platforms will allow for better assessment of vulnerable marine and aquatic habitats, as well as responses to hazards such as harmful algal blooms, oil spills, and storms that cause coastal flooding and erosion. Active sensors, such as Light Detection and Ranging (LIDAR), will allow us to probe into the depths of the oceans. Moreover, integrating surface ocean color measurements with three-dimensional measurements and models of the ocean will be increasingly important in discerning a changing ocean [49].

Finally, the approaches or algorithms for conducting ocean color remote sensing will be augmented as more spectral channels become routinely available and as ocean properties change. Purely statistical or empirical models are only accurate when conditions are similar to past conditions. When considering a changing ocean, the cause of the color



Remote Sensing of Ocean Color. Figure 8
Global climatologies or long-term averages of products derived from the Ocean Color SeaWiFS sensor from 1998–2011. (**a**) Chlorophyll *a* (mg m^{-3}); (**b**) colored dissolved organic matter (CDOM) index; (**c**) particulate inorganic carbon (PIC) (mol m^{-3})

change must be carefully assessed to separate the spectral variability due to phytoplankton from other sources of variability, such as sediments, CDOM, and even atmospheric aerosols. Considerable growth is also expected in approaches and technology for remote sensing of coastal habitats and assessing acute and chronic hazards. Comprehensive and consistent field observations from ships to autonomous vehicles and floats are required to assess the accuracy of satellite-derived products, build improved algorithms, and provide better linkages between surface measurements made from space and the processes within the water column [49]. Future effort will also be directed at assimilation of ocean color imagery into global circulation and climate models. As outlined above, remote sensing of ocean color is a complex discipline requiring radiometrically accurate and calibrated sensors, advanced techniques for atmospheric correction of aerosols and dust, and approaches that can deduce the source of variability in the color signal measured by a sensor. With the many important applications of ocean color remote sensing, from climate forecasting to environmental monitoring, a consistent and coordinated international investment in education, research, and technology is required to maintain and advance this dynamic field.

Bibliography

Primary Literature

1. Carr ME et al (2006) A comparison of global estimates of marine primary production from ocean color. Deep Sea Res Part II: Top Stud Oceanogr 53:741–770
2. Field CB, Behrenfeld MJ, Randerson JT, Falkowski P (1998) Primary production of the biosphere: integrating terrestrial and oceanic components. Science 281:237–240
3. Smith RC, Baker KS (1978) Optical classification of natural waters. Limnol Oceanogr 23:260–267
4. Martinez E, Antoine D, D'Ortenzio F, Gentili B (2009) Climate-driven basin-scale decadal oscillations of oceanic phytoplankton. Science 326:1253–1256
5. Siegel DA, Franz BA (2010) Century of phytoplankton change. Nature 466:569–570
6. Henson SA et al (2010) Detection of anthropogenic climate change in satellite records of ocean chlorophyll and productivity. Biogeosciences 7:621–640
7. IOCCG (2008) Why ocean colour? the societal benefits of ocean-colour technology. In: Platt T, Hoepffner N, Stuart V, Brown C (eds) Reports of the International Ocean-Colour Coordinating Group, Dartmouth, Canada
8. Morel A (1988) Optical modeling of the upper ocean in relation to its biogenous matter content (case I waters). J Geophys Res 93:10749–10768
9. Gordon HR, Morel AY (1983) Remote assessment of ocean color for interpretation of satellite visible imagery: a review. Springer, New York
10. O'Reilly JE, Maritorena S, Mitchell BG, Siegel DA (1998) Ocean color chlorophyll algorithms for SeaWiFS. J Geophys Res 103:24937–24953
11. McClain CR (2009) A decade of satellite ocean color observations*. Annu Rev Mar Sci 1:19–42
12. Kirk JTO (1994) Light and photosynthesis in aquatic ecosystems. Cambridge University Press, Cambridge
13. Blough NV, Del Vecchio R (2002) Chromophoric DOM in the coastal environment. In: Hansell DA, Carlson CA (eds) Biogeochemistry of marine dissolved organic matter. Academic, San Diego, pp 509–546
14. Twardowski MS, Boss E, Sullivan JM, Donaghay PL (2004) Modeling the spectral shape of absorption by chromophoric dissolved organic matter. Mar Chem 89:69–88
15. Ciotti AM, Cullen JJ, Lewis MR (2002) Assessment of the relationships between dominant cell size in natural phytoplankton communities and the spectral shape of the absorption coefficient. Limnol Oceanogr 47:404–417
16. Bricaud A, Claustre H, Ras J, Oubelkheir K (2004) Natural variability of phytoplanktonic absorption in oceanic waters: influence of the size structure of algal populations. J Geophys Res 109:C11010
17. Stramski D, Boss E, Bogucki D, Voss KJ (2004) The role of seawater constituents in light backscattering in the ocean. Prog Oceanogr 61:27–56
18. Mobley CD (1994) Light and water: radiative transfer in natural waters. Academic, San Diego
19. Gordon HR et al (2009) Spectra of particulate backscattering in natural waters. Opt Express 17:16192–16208
20. Twardowski MS, Lewis M, Barnard A, Zaneveld JRV (2005) In-water instrumentation and platforms for ocean color remote sensing applications. In: Miller R, Del-Castillo C, McKeee D (eds) Remote sensing of coastal aquatic waters. Springer, Dordrecht
21. Smith RC, Baker K (1978) The bio-optical state of ocean waters and remote sensing. Limnol Oceanogr 23:247–259
22. Morel A, Gentilli B (1993) Diffuse reflectance of oceanic waters. II. Bidirectional aspects. Appl Opt 32:6864–6879
23. Lee ZP, Carder KL, Arnone RA (2002) Deriving inherent optical properties from water color: a multiband quasi-analytical algorithm for optically deep water. Appl Opt 41: 5755–5772
24. Aurin DA (2010) Developing ocean color remote sensing algorithms for retrieving optical properties and biogeochemical parameters in the optically complex waters of Long Island Sound. Ph.D. Thesis. University of Connecticut
25. Ryan JP et al (2005) Coastal ocean physics and red tides: an example from Monterey Bay, California. Oceanography 18:246–255

26. Mouroulis P, Green RO, Wilson DW (2008) Optical design of a coastal ocean imaging spectrometer. Opt Express 16:9087–9096
27. Davis CO et al (2002) Ocean PHILLS hyperspectral imager: design, characterization, and calibration. Opt Express 10(4):210–221
28. McClain CR, Cleave ML, Feldman GC, Gregg WW (1998) Science quality SeaWiFS data for global biosphere research. Sea Technol 39:10–16
29. Gordon HR (1997) Atmospheric correction of ocean color imagery in the Earth Observing System era. J Geophys Res 102:17081–17106
30. Antoine D, Morel A (1999) A multiple scattering algorithm for atmospheric correction of remotely sensed ocean color (MERIS instrument): principle and implementation for atmospheres carrying various aerosols including absorbing ones. Int J Remote Sens 20:1875–1916
31. Gao BC, Montes MJ, Ahmad Z, Davis CO (2000) Atmospheric correction algorithm for hyperspectral remote sensing of ocean color from space. Appl Opt 39:887–896
32. Wang M, Son SH, Shi W (2009) Evaluation of MODIS SWIR and NIR-SWIR atmospheric correction algorithms using SeaBASS data. Remote Sens Environ 113:635–644
33. Yan B et al (2002) Pitfalls in atmospheric correction of ocean color imagery: how should aerosol optical properties be computed? Appl Opt 41:412–423
34. Fukushima H, Toratani M (1997) Asian dust aerosol: optical effect on satellite ocean color signal and a scheme of its correction. J Geophys Res 102:17119–17130
35. Antoine D, Nobileau D (2006) Recent increase of Saharan dust transport over the Mediterranean Sea, as revealed from ocean color satellite (SeaWiFS) observations. J Geophys Res 111: D12214
36. Claustre H et al (2002) Is desert dust making oligotrophic waters greener? Geophys Res lett 29:107–1
37. Paytan A et al (2009) Toxicity of atmospheric aerosols on marine phytoplankton. Proc Natl Acad Sci 106:4601
38. Garrison VH et al (2003) African and Asian dust: from desert soils to coral reefs. Bioscience 53:469–480
39. Monahan EC, O'Muircheartaigh I (1981) Improved statement of the relationship between surface wind speed and oceanic whitecap coverage as required for the interpretation of satellite data. In: Gower JFR (ed) Oceanography from space. Plenum, New York, pp 751–755
40. National Research Council Committee on Assessing Requirements for Sustained Ocean Color Research and Operations (2011) Assessing requirements for sustained ocean color research and operations. National Academies Press, Washington DC
41. Jerlov NG (1974) Optical aspects of oceanography. Academic, London, pp 77–94
42. Morel A, Prieur L (1977) Analysis of variations in ocean color. Limnol Oceanogr 22:709–721
43. Mobley CD, Stramski D, Bissett WP, Boss E (2004) Optical modeling of ocean water: is the case 1 – case 2 classification still useful? Oceanography 17:60–67
44. Morel A, Bricaud A (1981) Theoretical results concerning light absorption in a discrete medium and application to the specific absorption of phytoplankton. Deep-Sea Res 28:1357–1393
45. Siegel DA, Maritorena S, Nelson NB, Behrenfeld MJ (2005) Independence and interdependencies among global ocean color properties: reassessing the bio-optical assumption. J Geophys Res 110:C07011
46. Swan CM, Siegel DA, Nelson NB, Carlson CA, Nasir E (2009) Biogeochemical and hydrographic controls on chromophoric dissolved organic matter distribution in the Pacific Ocean. Deep Sea Res Part I: Oceanogr Res Pap 56:2175–2192
47. Dierssen HM (2010) Benthic ecology from space: optics and net primary production in seagrass and benthic algae across the Great Bahama Bank. Mar Ecol Progress Ser 411:1–15
48. Zaneveld JRV (1989) An asymptotic closure theory for irradiance in the sea and its inversion to obtain the inherent optical properties. Limnol Oceanogr 34:1442–1452
49. Dierssen HM (2010) Perspectives on empirical approaches for ocean color remote sensing of chlorophyll in a changing climate. Proc Natl Acad Sci 107:17073
50. Moore TS, Campbell JW, Dowell MD (2009) A class-based approach to characterizing and mapping the uncertainty of the MODIS ocean chlorophyll product. Remote Sens Environ 113:2424–2430
51. Schofield O et al (2004) Watercolors in the coastal zone: what can we see? Oceanography 17:25–31
52. Falkowski P et al (2000) The global carbon cycle: a test of our knowledge of earth as a system. Science 290:291
53. Behrenfeld MJ, Falkowski PG (1997) Consumers guide to phytoplankton primary productivity models. Limnol Oceanogr 42:1479–1491
54. Campbell J et al (2002) Comparison of algorithms for estimating ocean primary production from surface chlorophyll, temperature, and irradiance. Glob Biogeochem Cycle 16:1035
55. Westberry T, Behrenfeld MJ, Siegel DA, Boss E (2008) Carbon-based primary productivity modeling with vertically resolved photoacclimation. Glob Biogeochem Cycle 22:GB2024
56. Mouw CB, Yoder JA (2005) Primary production calculations in the Mid-Atlantic Bight, including effects of phytoplankton community size structure. Limnol oceanogr 50(4):1232–1243
57. IOCCG (2006) Remote sensing of inherent optical properties: fundamentals, tests of algorithms, and applications. In: Lee ZP (ed) Reports of the International Ocean-Colour Coordinating Group, Dartmouth
58. Siegel DA, Maritorena S, Nelson NB, Hansell DA, Lorenzi-Kayser M (2002) Global distribution and dynamics of colored dissolved and detrital organic materials. J Geophys Res 107:3228
59. U.S. National Aeronautics and Space Administration, Goddard Earth Sciences, Data and Information Services Center (2011) *Giovanni*. http://disc.sci.gsfc.nasa.gov/giovanni/

60. Behrenfeld MJ et al (2009) Satellite-detected fluorescence reveals global physiology of ocean phytoplankton. Biogeosciences 6:779–794
61. Dierssen HM, Kudela RM, Ryan JP, Zimmerman RC (2006) Red and black tides: quantitative analysis of water-leaving radiance and perceived color for phytoplankton, colored dissolved organic matter, and suspended sediments. Limnol oceanogr 51:2646–2659
62. Brewin RJW et al (2011) An intercomparison of bio-optical techniques for detecting dominant phytoplankton size class from satellite remote sensing. Remote Sens Environ 115:325–339
63. Balch WM, Kilpatrick KA, Trees CC (1996) The 1991 coccolithophore bloom in the central North Atlantic. 1. Optical properties and factors affecting their distribution. Limnol Oceanogr 41:1669–1683
64. Tomlinson MC, Wynne TT, Stumpf RP (2009) An evaluation of remote sensing techniques for enhanced detection of the toxic dinoflagellate, Karenia brevis. Remote Sens Environ 113:598–609
65. Simis SGH, Peters SWM, Gons HJ (2005) Remote sensing of the cyanobacterial pigment phycocyanin in turbid inland water. Limnol Oceanogr 50:237–245
66. Chavez FP, Ryan J, Lluch-Cota SE, Niquen CM (2003) From anchovies to sardines and back: multidecadal change in the Pacific Ocean. Science 299:217–221
67. Platt T, Csar Fuentes-Yaco KTF (2003) Marine ecology: spring algal bloom and larval fish survival. Nature 423:398–399
68. Fuentes-Yaco C, Koeller PA, Sathyendranath S, Platt T (2007) Shrimp (Pandalus borealis) growth and timing of the spring phytoplankton bloom on the Newfoundland–Labrador Shelf. Fish Oceanogr 16:116–129
69. Seibel BA, Dierssen HM (2003) Cascading trophic impacts of reduced biomass in the Ross Sea, Antarctica: just the tip of the iceberg? Biol Bull 205:93–97
70. Rosa R, Dierssen HM, Gonzalez L, Seibel BA (2008) Large-scale diversity patterns of cephalopods in the Atlantic open ocean and deep sea. Ecology 89:3449–3461
71. Dekker A et al (2005) Remote sensing of seagrass ecosystems: use of spaceborne and airborne sensors. In: Larkum AWD, Orth RJ, Duarte CM (eds) Seagrasses: biology, ecology, and conservation. Springer, Dordrecht, pp 347–359
72. Cavanaugh KC, Siegel DA, Kinlan BP, Reed DC (2010) Scaling giant kelp field measurements to regional scales using satellite observations. Mar Ecol Prog Ser 403:13–27
73. Phinn S, Roelfsema C, Dekker A, Brando V, Anstee J (2008) Mapping seagrass species, cover and biomass in shallow waters: an assessment of satellite multi-spectral and airborne hyper-spectral imaging systems in Moreton Bay (Australia). Remote sens Environ 112:3413–3425
74. Lesser MP, Mobley CD (2007) Bathymetry, water optical properties, and benthic classification of coral reefs using hyperspectral remote sensing imagery. Coral Reefs 26:819–829
75. Dierssen HM, Zimmerman RC, Drake LA, Burdige DJ (2009) Potential export of unattached benthic macroalgae to the deep sea through wind-driven Langmuir circulation. Geophys Res Lett 36:L04602
76. Burdige DJ, Hu X, Zimmerman RC (2010) The widespread occurrence of coupled carbonate dissolution/reprecipitation in surface sediments on the Bahamas Bank. Am J Sci 310(6):492–521. doi:10.2475/06.2010.03
77. Goes JI, Thoppil PG, Gomes HR, Fasullo JT (2005) Warming of the Eurasian landmass is making the Arabian Sea more productive. Science 308:545–547
78. Beman JM, Arrigo KR, Matson PA (2005) Agricultural runoff fuels large phytoplankton blooms in vulnerable areas of the ocean. Nature 434:211–214
79. Dwivedi RM, Solanki HU, Nayak SR, Gulati D, Somvanshi VS (2005) Exploration of fishery resources through integration of ocean colour with sea surface temperature: Indian experience. IJMS 34:430–440
80. Chavez FP, Strutton PG, McPhaden MJ (1998) Biological-physical coupling in the Central Equatorial Pacific during the onset of the 1997–98 El Nino. Geophys Res Lett 25:3543–3546
81. Lewis MR, Platt TC (1987) Remote observation of ocean colour for prediction of upper ocean heating rates. Adv Space Res 7:127–130
82. Hill VJ (2008) Impacts of chromophoric dissolved organic material on surface ocean heating in the Chukchi Sea. J Geophys Res 113:C07024
83. Gnanadesikan A, Anderson WG (2009) Ocean water clarity and the ocean general circulation in a coupled climate model. J Phys Oceanogr 39:314–332
84. Gnanadesikan A, Emanuel K, Vecchi GA, Anderson WG, Hallberg R (2010) How ocean color can steer Pacific tropical cyclones. Geophys Res Lett 37:L18802
85. Miller WL, Moran MA (1997) Interaction of photochemical and microbial processes in the degradation of refractory dissolved organic matter from a coastal marine environment. Limnol Oceanogr 42:1317–1324
86. Ackleson SG, Balch WM, Holligan PM (1994) Response of water-leaving radiance to particulate calcite and chlorophyll a concentrations: a model for Gulf of Maine coccolithophore blooms. J Geophys Res 99:7483–7499
87. Gordon HR et al (2001) Retrieval of coccolithophore calcite concentration from SeaWiFS imagery. Geophys Res Lett 28:1587–1590
88. Balch W, Drapeau D, Bowler B, Booth E (2007) Prediction of pelagic calcification rates using satellite measurements. Deep Sea Res Part II: Top Stud Oceanogr 54:478–495
89. Balch WM, Fabry VJ (2008) Ocean acidification: documenting its impact on calcifying phytoplankton at basin scales. Mar Ecol Prog Ser 373:239–247
90. Ryther JH (1969) Photosynthesis and fish production in the sea. Science 166:72–76
91. Wilson RW et al (2009) Contribution of fish to the marine inorganic carbon cycle. Science 323:359–362

92. Platt T, Sathyendranath S, Fuentes-Yaco C (2007) Biological oceanography and fisheries management: perspective after 10 years. ICES J Marine Sci 64:863
93. Platt T, Sathyendranath S (2008) Ecological indicators for the pelagic zone of the ocean from remote sensing. Remote Sens Environ 112:3426–3436
94. Stumpf RP et al (2003) Monitoring Karenia brevis blooms in the Gulf of Mexico using satellite ocean color imagery and other data. Harmful Algae 2:147–160
95. Hu C et al (2003) MODIS detects oil spills in Lake Maracaibo, Venezuela. Eos AGU Trans 84:313–319
96. Fingas M, Brown C (2000) A review of the status of advanced technologies for the detection of oil in and with ice. Spill Sci Technol Bull 6:295–302
97. Boland RC, Donohue MJ (2003) Marine debris accumulation in the nearshore marine habitat of the endangered Hawaiian monk seal, Monachus schauinslandi 1999–2001. Mar Pollut Bull 46:1385–1394
98. Donohue MJ, Boland RC, Sramek CM, Antonelis GA (2001) Derelict fishing gear in the Northwestern Hawaiian Islands: diving surveys and debris removal in 1999 confirm threat to coral reef ecosystems. Mar Pollut Bull 42:1301–1312
99. Randolph K et al (2008) Hyperspectral remote sensing of cyanobacteria in turbid productive water using optically active pigments, chlorophyll a and phycocyanin. Remote Sens Environ 112:4009–4019
100. IOCCG (2007) Ocean-colour data merging In: Gregg W (ed) Reports of the International Ocean-Colour Coordinating Group, Dartmouth
101. Roy S, Llewellyn C, Egeland ES, Johnsen G (2011) Phytoplankton pigments: updates on characterization, chemotaxonomy and applications in oceanography. Cambridge University Press. Cambridge Environmental Chemistry Series. Cambridge, UK. pp 845. ISBN: 978110700066-7

Books and Reviews

Campbell J, Antoine D, Armstrong R, Arrigo K, Balch W, Barber R, Behrenfeld M, Bidigare R, Bishop J, Carr ME et al (2002) Comparison of algorithms for estimating ocean primary production from surface chlorophyll, temperature, and irradiance. Glob Biogeochem Cycle 16:1035

Carr ME, Friedrichs MAM, Schmeltz M, Noguchi Aita M, Antoine D, Arrigo KR, Asanuma I, Aumont O, Barber R, Behrenfeld M et al (2006) A comparison of global estimates of marine primary production from ocean color. Deep Sea ResPart II: Top Stud Oceanogr 53:741–770

GlobCOLOUR: An EO based service supporting global ocean carbon cycle research. European Space Agency. http://www.globcolour.info/

IOCCG. Reports of the International Ocean-Colour Coordinating Group No. 1–10. Dartmouth. http://www.ioccg.org/reports_ioccg.html

Jerlov NG, Nielsen ES (eds) (1974) Optical aspects of oceanography. Academic, London

Miller R, Del-Castillo C, McKee BA (eds) (2005) Remote sensing of coastal aquatic waters. Springer, Dordrecht

Morel A (1991) Optics of marine particles and marine optics. In: Demers S (ed) Particle analysis in oceanography. Springer, Berlin, pp 141–188

Morel A, Bricaud A (1986) Inherent optical properties of algal cells including picoplankton: theoretical and experimental results. Can Bull Fish Aquat Sci 214:521–559

National Aeronautics and Space Administration (NASA) Ocean optics protocols for satellite ocean color sensor validation, vol I–VI. http://oceancolor.gsfc.nasa.gov/DOCS/

National Aeronautics and Space Administration (NASA) Ocean color web. http://oceancolor.gsfc.nasa.gov/

Platt T, Nayak S (eds) (2005). Special issue on: ocean colour remote sensing. Indian J Marine Sci 34(4):341–355

Siegel D (2004) Views of ocean processes from the sea-viewing wide field- of-view sensor mission: introduction to the first special issue. Deep Sea Res Part II Top Stud Oceanogr 51(1–3):1–3. http://dx.doi.org/10.1016/j.dsr2.2003.12.001

The Oceanography Society (2004) Special issue: coastal ocean optics and dynamics. Oceanography 17(2):1–95

Thomas A, Siegel D, Marra J (2004) Views of ocean processes from the sea-viewing wide field- of-view sensor (SeaWiFS) mission: introduction to the second special issue. Deep Sea Res Part II Top Stud Oceanogr 51(10–11):911–912. http://dx.doi.org/10.1016/j.dsr2.2004.06.003

Renewable Energy from Biomass, Introduction

MARTIN KALTSCHMITT
German Biomass Research Centre, Leipzig, Germany
Institute of Environmental Technology and Energy Economics, Hamburg University of Technology, Hamburg, Germany

Biomass has been a major energy source for humans since the mastering of the fire in ancient days. In addition, for a very long time in the history of humankind, bioenergy has been basically the one and only energy source for cooking, heating, and lighting. Beside this, biomass has been used as a horse or cattle feed for the provision of power and transportation duties and thus in an indirect way for energy provision. Therefore, the development of the human culture would have not been possible without the extensive use of biomass for energy.

During the last two centuries, the relative contribution of bioenergy within the overall energy system has been decreased due to an increasingly extensive use of, first, coal and, later, crude oil and natural gas. Nevertheless, on a global scale, biomass contributes still with roughly 10% to cover the overall primary energy consumption. Thus, bioenergy is still today the most important renewable source of energy used by humanity. It contributes in energy terms more within the global energy system as nuclear or hydropower. Moreover, this is true for developing as well as for industrialized countries. Biomass is not only the fuel for the poor people as it is often called in some developing countries.

In the years to come, the fossil fuel resources technically and economically easily accessible will decrease. Consequently, the prices for these commodities – and thus the overall energy price level – will increase due to the expected shortage resulting from the strongly growing energy demand especially from the emerging economies like China, India, and Brazil. Additionally, the international environmental politics will increase its efforts to reduce the greenhouse gas emissions for stabilizing the climate on a certain level (i.e., 2°C goal). Both developments will contribute to the fact that the importance of biomass for energy will continue to grow significantly in the future.

Biomass as a domestic energy carrier is stored solar energy; for this reason, it can be used depending on the actual given energy demand and can help to equalize the strong fluctuations of an energy provision from wind and solar likely to be used to a much higher extent in the years to come. Biomass is the most important carbon source available to humans beside the limited resources of fossil fuels (i.e., crude oil, natural gas, hard coal, and lignite). This is one of the reasons why additionally the combined provision of energy and raw materials (e.g., for the chemical industry) from biomass will gain more and more importance in the future (i.e., biorefinery approach). Moreover, biomass as an energy carrier can be used for the provision of heat and of electricity and for transportation purpose with existing technology available on the market ready for commercial applications. This makes organic material so valuable within a future energy system based compulsory to a much higher degree on renewable sources of energy compared to today for reasons outlined above.

Against this background, the goal of this section is to discuss various aspects associated with the energetic use of organic material. Therefore, first, biomass as a renewable source of energy is disputed in detail. Then, various possibilities to convert organic matter for the provision of heat, electricity, and fuels are presented under different frame conditions considering technical, economic, environmental, and social aspects.

To understand the possibilities and constraints of biomass for energy currently discussed very emotionally within politics and society, the *principles and prospects of biomass production* need to be understood from a biological point of view. Among other aspects, this makes it obvious how the overall efficiencies are between the solar radiation affecting the plant leaf and the energy content of the plant material synthesized from this solar energy. Based on this knowledge of the – from a biological point of view – possible plant productivity, the *worldwide biomass resources* are assessed to give an overview of the biomass potential available to humankind under various frame conditions. This shows that even if additionally to purely biological aspects other constraints (e.g., water and nutrient availability, appropriate production schemes, ecological criteria) are taken into consideration (which has to be the case without any doubts), the biomass potential available for the energy markets without affecting the markets for food and fodder and for raw materials is still huge. This is true even against the background of the enormous global primary energy consumption as well as the potentials of other renewable sources of energy. Nevertheless, due to the food vs. fuel discussion, an intensive search for "new" or alternative biomass resources, which cannot be used within the food and fodder markets, is ongoing, promising much higher area specific yields and/or a much better use of the solar insulation. This is especially true for the assessment of the *possibilities and constraints of algae as a "new" biomass resource* discussed currently as the big white hope especially for the provision of fuels for transportation purpose.

However, the biomass potential available to fuel the world population is already used to certain extents. Although it is impossible for fundamental reasons to assess accurately the overall current use (e.g., due to insufficient data), the *biomass use on a global scale* can be roughly estimated based on robust assumptions and approximate calculations. This allows making valuations on how the use of biomass as an energy carrier can be expanded on a global scale. Additionally, it shows very clearly that biomass contributes today to all important energy markets (i.e., heat, electricity, fuels). Beside this, it becomes obvious that the contribution of biomass within the global energy system is strongly increasing in absolute terms. This is especially true for a modern use of biomass realized in industrialized countries and with an increasing share also in developing nations. Additionally, biomass is used in a "traditional" way in developing countries mainly for the provision of energy for cooking and heating. However, the assessment of the potentials and the use shows that partly biomass use exceeds production. This means that a nonsustainable situation is given in some parts of this world, as it is, for example, the case in some densely populated and very poor areas in selected developing countries. Therefore, *sustainability aspects of a biomass provision and use* are gaining more and more attention in energy politics and economy on a global as well as on a national scale. Consequently, some governments have implemented already quite strict sustainability criteria to be fulfilled if biomass-derived fuels should be used within the energy system especially under the viewpoint of climate protection. Thus, the production, provision, and use of biomass as a sustainable energy carrier (i.e., fulfilling the given sustainability demands) become more and more complex and demanding. Therefore, *possible conversion routes for biomass as an energy source* are analyzed and discussed to give an idea about the huge variety of the given possibilities.

Based on this overview of various provision routes and their determinants, different biomass resources available for energy purpose are discussed in detail. First of all, this is true for the *production and provision of woody energy crops* from "classical" forests as well as from short rotation plantations (i.e., agricultural production of wood). These wood resources are usually used for the provision of solid fuels. Based on this, *solid biofuels and their characteristics* as well as *upgraded new solid biofuels (like pellets, torrefied wood, etc.)* are presented in detail. This discussion makes it obvious that wood logs and wood chips are still dominating the local wood fuel markets. However, wood pellets as a globally traded commodity gain more and more market importance in recent years, and this development is likely to continue in the years to come. Beside these wood-based biofuels used mainly in combustion devices, substrates for biogas production are used in an increasing manner in recent years among others due to effective subsidy schemes in industrialized countries as well as in emerging economies. Therefore, the *provision of biogas substrates from municipalities and from industry* as well as *production and provision of energy crops for biogas production* is tackled in detail, showing the fuel characteristics as well as the production and provision options including the determining parameters.

The biomass resources discussed above are used traditionally to a large extent within the heat market (i.e., for the provision of thermal energy mainly for cooking and space heating). This application field can be subdivided in several subareas. Depending on the thermal capacity installed within the combustion device, *heat provision in modern small-scale systems* and *heat provision in modern large-scale systems* can be distinguished. Both options are characterized by very advanced conversion technologies, minimizing the environmental burden and maximizing the overall conversion efficiency between the fuel energy and the provided useful energy. Thus, it becomes obvious that today combustion of solid biofuels is realized in high-tech devices, taking the latest findings from numerous research activities into consideration. The fact that one can smell if the neighbor is heating his house based on wood fuels is history. Additionally, *small-scale heat and power (CHP) systems* are gaining more and more importance for the provision of heat and power to cover the local demand of both energy carrier and to maximize the exergetic conversion efficiencies in parallel. However, most of these small-scale CHP options are still in a developing phase. They have gained only very limited market relevance yet. For developing countries with a lack of access to modern, high-tech conversion devices for solid biofuels, *heat provision for cooking and heating* is realized still with traditional technologies (like the three-stone oven) characterized by rather low conversion efficiencies as well as high environmental impacts especially of emissions with a toxic impact on humans. One option to overcome the manifold problems resulting from the use of solid biofuels (e.g., dried dung) in such primitive devices in these poor regions could be *biogas production in developing countries for small-scale heat and cooking applications.* This possibility is characterized by various advantages if it is implemented locally in an intelligent way (which is unfortunately often not the case in reality).

During the last century, more and more biomass is used for electricity generation, helping to cover the strongly increasing demand for electrical energy in developing as well as in industrialized countries. Therefore, the various options converting organic material into electricity to be fed into the local and supraregional electricity grids are discussed in detail. This is especially true for the "classical" and most widely used option – *biomass combustion for electricity generation.* Within such combustion devices coupled to a conventional power plant cycle or maybe an Organic Rankine Cycle (ORC), basically, only solid biofuels – like wood chips, wood logs, wood pellets, and sometimes even straw – are used with installed electrical capacities varying between a couple of megawatts up to some 100 MW. Especially in countries with cold winter months and thus a high space heating demand, these plants are often operated in a CHP mode to enhance the overall efficiency and thus to reduce the costs of these systems. Beside such an electrification or combined heat and power generation based exclusively on solid biofuels, *biomass cofiring in large-scale coal-fired power plants* is realized with an increasing importance. This option of burning solid biofuels together with coal in existing power plants allows to reduce the investment costs, to increase the overall conversion efficiencies, and to contribute to greenhouse gas mitigation. Due to these reasons, this option has gained more and more importance for electricity generation partly coupled with heat provision especially in industrialized countries.

Nevertheless, due to an increasing lack of cheap solid biomass on the local and global markets, the overall development aim is to provide electricity from such feedstock with a much higher overall electrical efficiency resp. an increased exergetic conversion rate. The key technology promising to allow for significantly improved efficiencies compared to the "classical" combustion with a steam cycle coupled to a combustion device is the gasification of solid biofuels to a gaseous energy carrier to be used in an engine, a gas turbine, a fuel cell, or even an IGCC (i.e., integrated gasification combined cycle). Therefore, the *large-scale biomass gasification for the provision of electricity and fuels* is discussed, showing that the technology is on the borderline to market introduction even due the fact that some technological challenges have to be met in the future. Additionally, the *small-scale biomass gasification for rural electrification* allows in principle to provide electricity with high efficiencies for small-scale, local, and decentralized applications (e.g., electricity generation for stand-alone application). Therefore, this option is tackled in detail, showing that the technology is fundamentally available but much too expensive for the time being. Nevertheless, it becomes obvious that on the shorter or longer term, gasification will play a much more important role within the conversion of solid biofuels into electricity due to fundamental thermodynamic and technological advantages compared to combustion.

Beside solid biofuels dominating the markets for biomass feedstock for electricity generation so far, also liquid and gaseous bioenergy carrier are more and more widely used for the conversion to electrical energy partly coupled with the provision of thermal energy. Therefore, the *possibilities and constraints of liquid biofuels for combined heat and power* (*CHP*) as well as *biogas provision* (*high-tech applications*) for electricity generation are discussed in detail. These two contributions show that for certain niche markets and/or specific applications, these options are very promising according to technical, economic, and especially environmental aspects.

In accordance with the development of the markets for heat and electricity, also the use of biomass-derived liquid and gaseous fuels to be used within the transportation sector has gained more and more market relevance in recent years. For example, liquid biofuels (i.e., bioethanol and biodiesel) contributed in 2010 with 0.5% to cover the overall global primary energy consumption. One main driver for this development is the fact that biomass is the one and only renewable source of energy which can be used with available and market-ready technology within the existing transportation system. For these reasons, the production and provision of *biodiesel* is discussed, showing that the processes to provide a fuel with similar characteristics compared to fossil diesel fuel are rather simple.

Therefore, such conversion processes can be implemented also in rural areas of developing countries; this is one reason why countries with a strong vegetable oil production industry are promoting this provision route with administrative measures. The same is true – maybe to a minor extent – also for bioethanol. However, the globally existing ethanol production can be subdivided in roughly two different markets representing two slightly varying technologies based on two differing types of biomass feedstock: bioethanol from sugar and from starch. The former is realized primarily in Brazil (and other tropical countries) and the latter in the US (and other countries in temperate zones). Both countries provide together more than 85% of the globally produced bioethanol. To cover these two different provision chains in an adequate way, firstly, the production and provision of *bioethanol from sugar – the Brazilian experience –* is discussed in detail. Secondly, aspects related to *principles and available experiences for bioethanol manufacturing from starch* are presented. Both essays as well as the development of the global bioethanol markets show clearly that the technology is market ready. Nevertheless, there are manifold ways to improve the processes to maximize the overall conversion efficiency and to minimize the waste products. Anyhow, due to fundamental biological and physical reasons, the overall conversion efficiency from biomass to bioethanol is limited, and the overall energy demand for the production of dehydrated ethanol ready to be used within a flex-fuel vehicle is high in bioethanol factories realized today. Therefore, globally enormous research money is currently spent to develop better and more integrated processes (i.e., biorefinery approach) and to expand the resource basis to increase the hectare-specific bioethanol yield. One option for the latter is the production of *bioethanol from celluloses*. Therefore, this option is discussed in detail with a special focus on the presentation of the ongoing developments of principles and processes for a most efficient use of lignocellulosic biomass.

Such processes for the conversion of lignocellulosic biomass via fermentation to bioethanol compete with other options for the provision of liquid and gaseous fuel based on solid biomass. On the one hand side, this is true for the thermochemical conversion of biomass, e.g., Fischer–Tropsch diesel, to be used as a liquid fuel in existing diesel engines. Therefore, *biomass to liquid* (*BtL*)concepts are presented and assessed, showing that based on a thermochemical gasification of solid biomass and an upgrading of the provided gas, basically all fuels currently used within the transportation sector can be provided. The same is true for *biomass pyrolysis* promising also the provision of liquid biofuels for a possible use as a transportation fuel. On the other hand, also gaseous fuels can be provided based on thermochemical processes, allowing for relatively high conversion efficiencies. For these reasons, the provision chains for *bio-synthetic natural gas* (*Bio-SNG*) as well as *biohydrogen* are outlined and discussed in detail. Additionally, *biomethane from anaerobic processes* can also be used as a substitute for natural gas in CNG vehicles. Thus, the experiences and perspectives available so far are also debated in detail for this biochemical conversion route. Finally, a *technical, economic, and environmental comparison of the different biofuels* is carried out, indicating that there is no silver bullet for the provision of biofuels available so far.

All over, this section gives a broad overview on biomass as an energy carrier for the provision of heat, electricity, and transportation fuel considering technical, economic, environmental, and social aspects.

Renewable Generation, Integration of

Bri-Mathias Hodge[1], Erik Ela[1], Paul Denholm[2]
[1]Transmission and Grid Integration Group, National Renewable Energy Laboratory, Golden, CO, USA
[2]Energy Forecasting and Modeling Group, National Renewable Energy Laboratory, Golden, CO, USA

Article Outline

Glossary
Definition of the Subject and Its Importance
Introduction
Electricity System Background
Characteristics of Renewable Resources
Generator Modeling and Interconnection

Operating Impacts
Market Considerations
Transmission Planning
Enabling Greater Renewable Penetration
Future Directions
Bibliography

Glossary

Ancillary services All of the actions necessary for supporting the transmission of power from the generator to the consumer and ensuring reliable system operations. Some examples of these services include: voltage and frequency control, generation scheduling, load following, and system protection.

Balancing area An area in which electricity supply and demand are locally matched and over which a balancing authority maintains system frequency and provides operating reserve.

Independent system operator The organization that is charged with controlling the operation of the electrical power transmission system in a certain geographic area.

Operating reserve Extra generating capacity available at short notice to replace scheduled capacity that is currently unavailable due to some sort of system disruption.

Unit commitment and economic dispatch The process by which generators are scheduled by the grid operator in order to meet expected demand at all timeframes. Commitment refers to deciding which generators will be turned on far in advance of the time period under consideration. Dispatch refers to the decision of how much power each generator will supply during a timeframe that is closer to realization than the commitment period.

Variable generation Generation from units that cannot be well controlled and thus are not perfectly dispatchable. This term is often applied to generation from weather-driven units, such as wind and solar.

Definition of the Subject and Its Importance

The integration of renewable generation consists of all of the changes in power system operations that are required in order to allow renewable generation sources to play a significant role in the electricity system. The impacts are mostly due to variable generation (VG), like wind and solar power. Historically these technologies have been labeled as intermittent generation, but recent trends prefer the label variable generation [1]. Variable generators have a maximum available generation limit that changes with time (variability) and this limit is not known with perfect accuracy (uncertainty). This uncertainty and variability is in addition to that of the existing system and can therefore create additional challenges for grid operators to maintain their current levels of reliability.

Introduction

Renewable electricity generation encompasses a number of distinct technology types, often classified by their power source, such as geothermal, hydroelectric, marine, biomass, solar, and wind power. Within each of these groupings there are a number of different technologies for harnessing the energy of the power source. However, when discussing renewable integration wind, marine, run of river hydro, and solar generation tend to garner most of the interest. This is due to the fact that the output of these plants is variable and uncertain. Other types of renewable generation are more similar to traditional power sources, such as fossil fuels and nuclear power, in that they are dispatchable. This means that they can be reliably scheduled in advance to provide power when desired, and do not need to rely on the current weather conditions. This is the result of the fact that the availability of their energy source can be controlled. While no generator can guarantee availability at a scheduled time, every generator has the possibility of being unavailable. The unplanned outage rates for dispatchable generators are low enough for the system to treat them as if they will produce at the desired level in the scheduled timeframe, while holding contingency reserves should an outage occur. While other renewable generating units can occasionally have uncertain output, for example, hydroelectric units cannot operate below certain reservoir levels, they are normally treated similarly to conventional units. For this reason the integration of these generating technologies is not normally considered an issue for electrical system operation. Another way of demonstrating the difference between variable generation and dispatchable generation is to examine some of the factors used to describe their patterns of usage. One common metrics is the capacity factor. A capacity factor is the amount of electricity a unit would be physically able to generate divided by the theoretical maximum that the unit would produce if it ran at full capacity over a certain time period. Therefore times when the unit would be down or at a reduced capacity due to forced and unforced outages count against the capacity factor. Baseload power plants typically have capacity factors on the order of 90% or higher. Wind power plants sited in good onshore locations have capacity factors of around 30%, while solar PV plants, even in very good locations, tend to have capacity factors under 20%. However, capacity factor alone does not tell the whole story. For example, a natural gas turbine that is used only for peaking may have a capacity factor under 5%. Even though the turbine is available for a greater percentage of the time, it is not always chosen in the unit commitment and dispatch process due its higher operating costs than baseload plants. For this reason other metrics such as forced and unforced outage rates are also used to characterize unit usage. These metrics however do not apply as well to variable generators and therefore metrics such as capacity value and effective load-carrying capability are often utilized when discussing the availability of wind and solar generators. When the integration of renewable generation is considered, wind and solar generation are usually the focus due to their variable and uncertain nature and their current presence in relatively large quantities in the system.

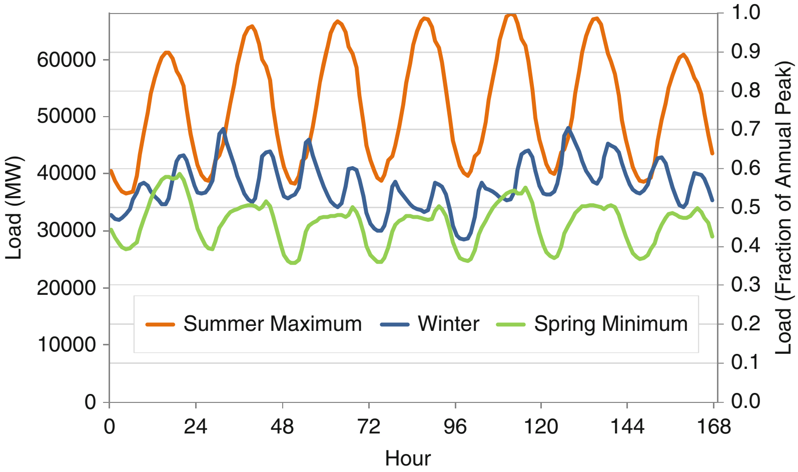
Worldwide wind power output grew sevenfold and solar photovoltaic (PV) production grew 16-fold from 2000 to 2008 [2]. Wind power has been the fastest growing source of electrical generation capacity in the United States for the last several years [3]. In 2009, there were over 34 GW of wind capacity operating in the United States and over 600 MW of solar PV [3]. These trends have transformed variable generation from a minor component in the electricity system to a contributor whose effects on the overall system must be thoroughly considered. As the future electricity system is expected to contain even larger quantities of variable generation, it is important that means of integrating variable generation are well researched before the higher penetration rates are achieved.

Electricity System Background

The current electricity system is the product of over 100 years of evolution and growth. A central pillar of the system is dispatchable generation, generally from fossil fuels, hydroelectricity, and nuclear power. As most large-scale power systems have relatively small amounts of electricity storage, electrical supply must always meet electrical demand. The instantaneous matching of supply and demand is needed in order to maintain a nominal electrical frequency. When supply does not equal demand, the frequency can change from its scheduled value (60 Hz in the North America, 50 Hz in many other locations). Frequency that deviates too far from its scheduled value can trigger underfrequency load shedding or over-frequency relays disconnecting machines to prevent damage. If this is not controlled, cascading events can lead to blackouts. This need for supply–demand balance coupled with the fact that the vast majority of demand is noncontrollable necessitates the current structure of system operation. Electricity system operation is a complex process where demand must be forecast and generation scheduled in advance, but numerous types of reserves must also be kept waiting in order to ensure system reliability, should forecast demand errors occur or scheduled generation units become unavailable.

Electricity demand follows strong seasonal and daily patterns. Seasonal demand patterns are highly correlated with seasonal weather patterns. Most systems in the United States tend to have their highest demand during the summer due to the large additional loads attributable to air conditioning. European systems generally tend to reach peak levels during the winter due to the demand from electric space heating. Figure 1 demonstrates both the seasonal and daily patterns that occur in the Electric Reliability Council of Texas (ERCOT) system. As may be observed, the summer peaks can be as much as double the spring minimum loads. The daily differences can also be very large during the summer months. Figure 1 shows an instance where total system load varies between 40 and 60 GW in the course of a single day. In order to meet these vastly different conditions, utilities build generating units with very different production characteristics.

Baseload power plants are utilized to cover the large amounts of demand for electricity that fall below the

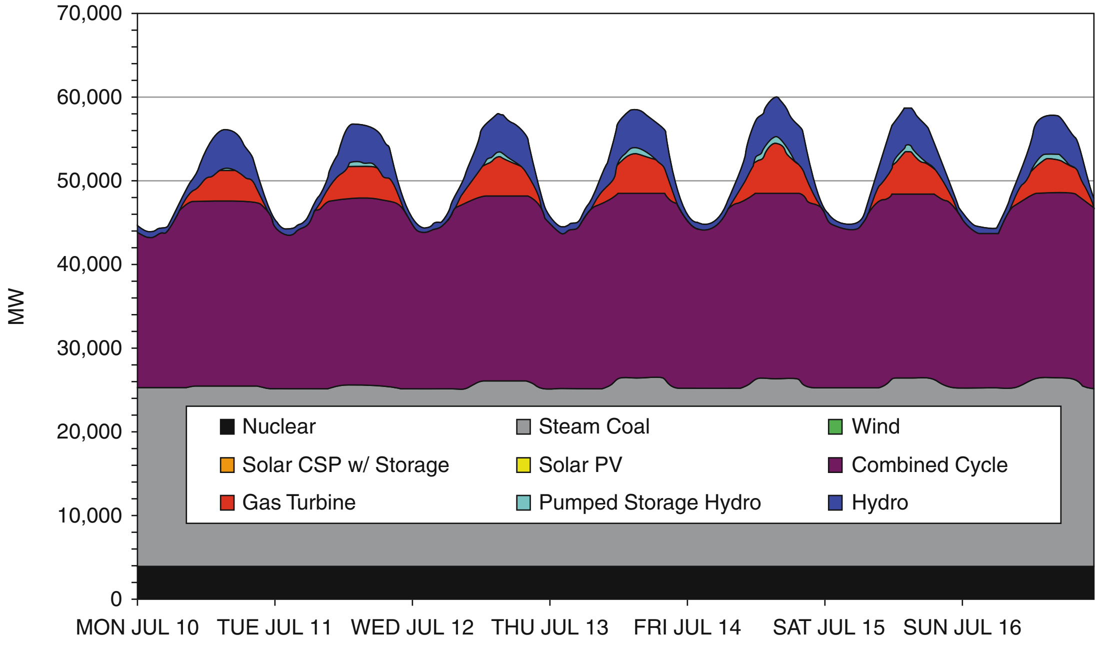


Renewable Generation, Integration of. Figure 1
Hourly loads from ERCOT 2005 [4]

minimum daily demand level. Coal-fired and nuclear plants are the two most typical types of baseload plants. They are often very high capacity plants that have very large capital costs, and relatively low operating costs. Baseload plants also typically require long start-up times and cannot ramp rapidly to follow changes in load. For these reasons utilities prefer to run the plants as close to maximum output as possible the majority of the time they are in service. Variations in load are usually met through the use of another type of generating unit, the load following plant. These units can be further subdivided into intermediate load plants and peaking plants. Intermediate plants are typically used to meet the daily variability in demand and are often hydroelectric or combined cycle plants fueled by natural gas. Peaking plants usually have a very high marginal cost of production, and as such are only utilized during periods of extremely high demand, often less than a few hundred hours per year. These plants are often natural gas or oil-fired simple cycle turbines. The normal usage patterns of baseload, intermediate, and peaking units can be observed in Fig. 2.

The generating units that will meet the expected load are usually scheduled 1 day beforehand in what is known as a security-constrained unit commitment. This optimization of total system costs schedules the start-up of units based on forecasted loads for the next day. Because system load cannot be perfectly forecast, there is the need for another assignment process closer to the actual time point. This process is known as security-constrained economic dispatch and it changes the level of output of units already online in order to ensure that supply meets demand. The amount of time ahead that this process occurs varies by system operator. Advances in telecommunications have allowed some systems to operate at smaller dispatch timings, down to 5 min ahead. Both of these processes should ensure that the commitment and dispatch are secure considering the generator constraints, transmission network constraints, and n-1 contingency constraints. The n-1 criterion states that a system must be secure following any credible single outage, be that a generator, or other system component (e.g., a transmission line).

In addition to advanced scheduling that needs to occur to balance the generation, system operators often schedule capacity as operating reserve to be used in case of unexpected changes from schedules or for variability occurring within a scheduling interval. Operating reserves can be classified according to the type of situation that triggers their deployment, the timescale of the response, and the direction of deployment. Figure 3 illustrates the different types of reserves based upon the event type and the speed of response. The most commonly utilized reserves are those that are required during normal operating conditions. Generating units must be kept in reserve in order to respond to continuous differences between forecasted and actual conditions. This function falls under the category of reserves



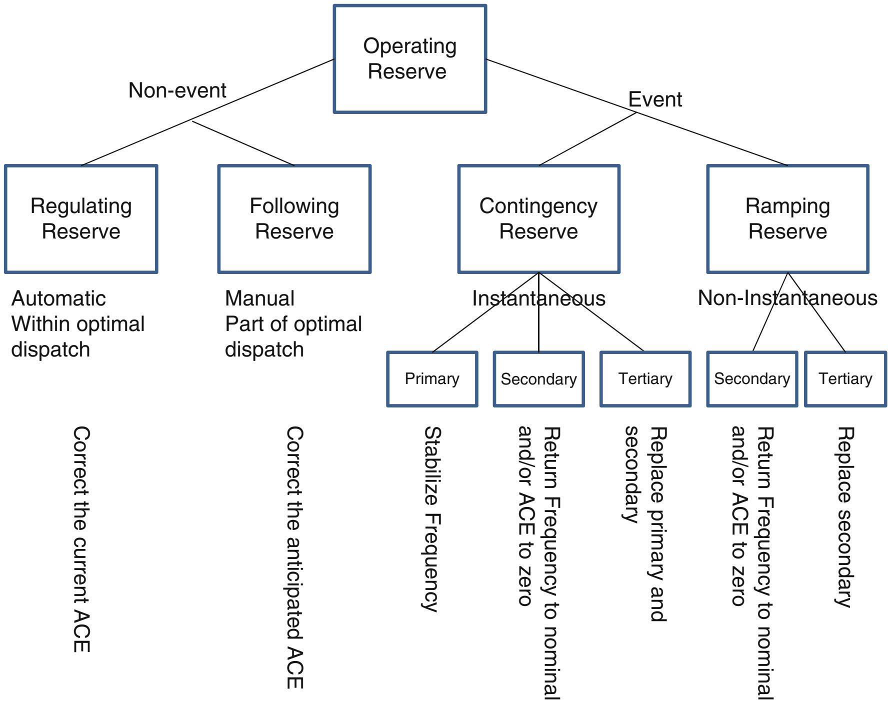
Renewable Generation, Integration of. Figure 2
Normal unit commitment and dispatch patterns in the western United States with a very low renewable generation penetration rate [5]

known as regulation reserve. It is employed to respond to the minor random fluctuations that occur around normal load in order to maintain system frequency. Regulation reserves are required in both the up and down directions, in other words, regulation reserve must be able to both increase and decrease output to match the fluctuating conditions. Regulation reserves are generally employed on the sub-minute timescale. More sustained trends over the timescale of minutes are handled through following reserve (often called load following reserve in practice).

In addition, reserves must be kept on hand should a generating plant or transmission line currently importing power suddenly become unavailable. Contingency reserves are those in place for unexpected events that change the instantaneous availability of generators or transmission. Primary reserve, that is, frequency responsive reserves, responds within seconds in order to stabilize the system frequency after a major disturbance. Secondary reserve is then used to restore the frequency back to its scheduled setting. Tertiary reserve is then used to replace the reserve used during the event so that the system is secure toward a subsequent event. Ramping reserve is utilized in order to respond to events that occur over a longer duration, such as variable generation forecast errors or ramping events. Ramping reserve also requires frequency or control error to return to its nominal setting and for it to be replaced in case of a secondary event. However, due to the slow time it takes for these events to occur it is usually not necessary to stabilize the quickly changing frequency. Reserves are also often classified according to their synchronization status. Reserves provided by units that are already running and synchronized to the system are known as spinning reserves, while non-spinning reserves are not synchronized and thus take longer to respond. Fast response reserve, such as regulating reserve and primary or frequency response reserve, require spinning reserve exclusively while slower reserve categories usually contain mixtures of both spinning and non-spinning reserves in different proportions.

Characteristics of Renewable Resources

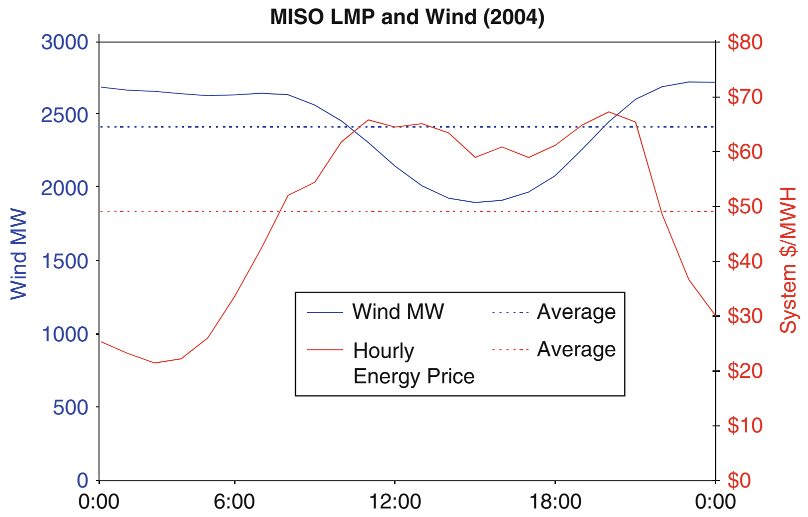
As described in the Introduction, the discussion of renewable resource characteristics is generally limited



Renewable Generation, Integration of. Figure 3
Illustration of reserve types [6]

to wind and solar generation, since they operate quite differently from other types of generation. While both are considered variable resources, wind and solar generation have distinct characteristics that need to be considered when attempting to integrate them into the electricity system. Perhaps the most important difference is in their diurnal patterns. Solar generation output follows the daily cycle of the sun in its particular location on earth, and thus is limited to daylight hours. Wind power output may occur during both daylight and night hours, but in most locations output has a tendency to be higher and more consistent during the night hours. This is demonstrated in the plot of average wind power output over the course of a day for the year 2004 in the MISO system seen in Fig. 4. Plotted on the same graph are the average locational marginal prices (LMPs) for the MISO system over the course of a day. As may be observed, the wind output tends to be highest when energy prices are lowest, corresponding to times of low demand, and dips during the daily peak when prices are highest.

Solar generation can be subdivided into two different categories: concentrating solar thermal and solar PV. From an operational perspective the biggest difference between the two is that concentrating solar thermal plants can have thermal storage capabilities. This allows them to store energy that can later be used to provide electricity during periods where output would be diminished if relying solely on immediate solar irradiation. Solar PV on the other hand is completely reliant on the immediate solar irradiation and thus has reduced output in cloudy weather, and no output at night. The difference in daily profiles for concentrating solar thermal plants with storage versus PV plants can be seen in the July portion of Fig. 5. Another major difference between the two technologies is the locations in which they can be deployed. Concentrating solar thermal plants require larger areas for installation and the high capital costs dictate that they be deployed in only very high quality resource areas with high direct normal irradiance. Solar energy contains both a direct and a diffuse component. The direct component is the light from the solar beam and the diffuse component consists of light that has



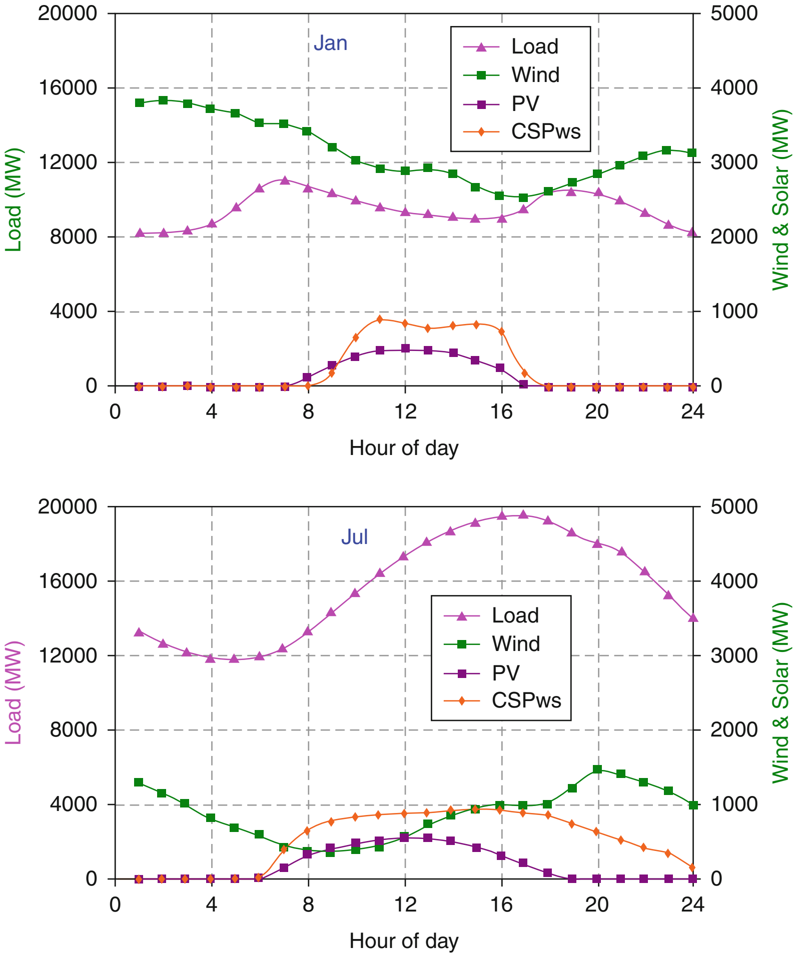
Renewable Generation, Integration of. Figure 4

Average energy prices and wind power output over the course of the day in the MISO system [7]

been scattered by the atmosphere. The direct component can be focused using mirrors or lenses and accounts for 60–80% of total solar insolation in clear sky conditions, but decreases with high levels of humidity, cloud cover, and atmospheric aerosols. This limits their construction to arid regions, such as those in the southwestern United States. PV on the other hand can be deployed in nonutility scale system sizes and thus can be present in large plant configurations or distributed over a number of locations, such as rooftops. This can be an important distinction in the case of very high levels of distributed PV penetration as the output from these locations only appears to the system operator as reduced load.

Wind generation is also subdivided into two categories, though usually based on the location of the wind plant, not the technology being used. Onshore locations have been the dominant choice thus far as wind power capacity has grown. However, offshore locations generally have more consistent output and the accompanying higher capacity factors are attractive. Offshore locations tend to involve the construction of larger turbines, and are usually more expensive due to the difficulties associated with their construction in large bodies of water. While onshore locations can require building new transmission, the additional transmission demands of offshore locations are more extensive, but conversely have the advantage that they are generally located closer to load centers.

One very important positive aspect of both wind and solar is the large size of the resource base. For example, while global electricity consumption in 2005 was estimated to be 16.6 PWh, the resource base of global onshore wind power, restricted to locations with a minimum 20% turbine capacity factor, is approximately 700 PWh [8]. The global offshore resource base, while not as large, is still considerable. Even after restricting the locations to those within 50 nautical miles of the coast and water depth of 50 m or less, the potential energy production is still approximately 80 PWh [8]. The total potential capacity of offshore wind power in the United States is over 4,000 GW, based on locations within 50 nautical miles of the coast with average annual wind speeds at 90 m greater than 7 m/s [9]. Solar PV power has an even larger potential resource base, with the technical potential estimated to be approximately 4,100 PWh per year by 2050 [10], while concentrating solar thermal potential is estimated to be between 630 and 4,700 GW of capacity [11]. As a point of comparison, small and micro hydroelectric plants have an estimated global technical potential of 150–200 GW of capacity [11]. These enormous resource bases, combined with the current low utilization of the potential energy from these sources, make renewable resources attractive options for long-term planning of energy supply sources.



Renewable Generation, Integration of. Figure 5

Load, wind, solar pv, and concentrating solar thermal (with storage) daily profiles for both January and July in Arizona in a high renewable penetration scenario [5]

Because these variable generation sources are not dispatchable, they present new issues for integration into the electricity grid. One major difference is that they are often considered to be a negative load, instead of a generation source. Essentially, any variable generation that is being produced at the current time is accepted and reduces the total load that must be met by conventional generators. This residual load is equal to normal demand minus electricity from renewable generators, and has greater variability than load alone. While this is a suitable representation for low levels of variable generation penetration, it will need to change in order to accommodate higher levels. In fact this is already starting to change as wind is being used to alleviate transmission constraints.

Generator Modeling and Interconnection

There are many technical aspects that concern the physical connection of renewable plants to the electricity grid, and their contributions to stable system operation. As renewable power becomes a larger contributor to the electricity system, it must also take on the responsibilities of maintaining system security. We focus mostly on wind in this section due to the currently higher market share. The requirements of wind turbines connected to the electricity system are prescribed through grid codes that detail the turbine contributions to ensure power system stability. The interactions that occur between the grid and the wind plant are heavily influenced by the type of machine in use and the stiffness of the grid. There are four different types of wind turbine machines commonly in use (Types 1–4), differing in their ability to control output power and reactive power and the type of generator utilized. The creation of accurate models of the physical representation of these different turbines can aid in the understanding of the effects that wind plant interconnection will have on the system. Small wind plants may be connected at low or medium voltage distribution networks. In a larger plant many turbines are connected together through a collector system and then connect to the substation where a transformer steps up the voltage from the distribution level to the transmission level.

Fault Ride Through

One common requirement made of wind turbines is fault ride through in the case of a system fault. Early wind plants would simply disconnect from the system when a system fault occurred. In systems with larger amounts of wind power this disconnection would exacerbate the problem [12]. This requirement specifies that the generator must stay online during faults with a duration and voltage variation under certain thresholds that vary by system. Further advances in wind turbine technology will allow for the post-fault recovery characteristics of turbines to be better than those of conventional generators. While wind turbines must comply with the requirements for normal network operation, they must also have system protection that prevents damage to the turbine during events that take the system out of its normal operating ranges. Plants typically are required to have protection from over- and under-frequency, current and voltage events. Additionally, the type of power electronics utilized by any renewable generation can significantly enhance its response to events.

Frequency Control

Frequency fluctuations in the power system result from an instantaneous difference between the amount of power being generated and the demand, including system losses. Wind plants can contribute to the stabilization of frequency in over-frequency events by decreasing output, either through blade feathering or shutting down individual turbines. Under-frequency events are more challenging because they require additional generating capacity. Since the wind speed cannot be increased to produce more generation when desired, this requires holding a portion of possibly generating capacity free, despite the lost revenue incurred. If decreases in the wind power output can be forecast, the turbines can reduce their output slowly in advance reducing the impact of the negative ramp rate.

Voltage Control

In order for a power system to operate as designed, the voltage throughout the system should be kept within the normal operating range (0.95–1.05 of the set point), though transients are allowed a wider range (0.9–1.1). The impedance of a power line causes a change in voltage between the two ends when a current is flowing. The introduction of renewable generators changes the power flow in the system, and hence the local voltages. Unlike frequency, voltage is a local phenomenon that cannot be controlled system-wide. Thus voltage must be kept within its specified range locally by generators in the area, or by power electronics, such as tap-changing transformers. Voltage control in transmission lines is normally adjusted through the consumption or supply of reactive power by generators, or through capacitor banks or flexible AC transmission system (FACTS) devices.

Reactive Power

Wind turbines with induction generators consume reactive power, which can lead to voltage collapse, if not properly compensated, and increased system losses from voltage drops and reactive current, respectively. To mitigate the effects, reactive power can be supplied through the inverter on variable speed turbines, or through capacitors in fixed speed turbines.

Interconnection Queue Process

The process by which a new generator receives the requisite permissions from the local grid operator to connect into the transmission system is known as the interconnection queue process. This process enables the system operator to conduct feasibility, system impact, and facility studies to ensure that the new generator will not have negative effects on system operation. The process also includes the planning steps necessary for generator participation in local markets and the acquisition of necessary transmission rights.

Operating Impacts

With the increasing amounts of wind power being added into electricity systems in recent years, a number of studies have been conducted in order to assess the impacts on the system incurred by integrating these large amounts of wind power. These studies typically examine two cases, one with wind and one without, and compare the results of statistical and production cost simulation analysis in order to determine the cost differences in system operation that may be attributed to wind power. A common assumption is that system reliability must be held at a constant level, which may necessitate the inclusion of additional operating reserves in the with wind case to remedy the increased system variability and uncertainty introduced by the wind plants.

Regulation

The goal of regulation is to compensate for the system variability at small timescales. The majority of this variability is due to normal load fluctuations, with an additional component of variability attributable to conventional generators minor deviations from their set points. The addition of renewable generation to the system adds another source of variability that must be accounted for. However, the impacts of wind generation on regulation services have been found to be relatively minor. For example, the addition of 3300 MW of wind into a system with approximately 30 GW of system peak load required only 36 MW of additional regulation [13]. This is due to the fact that the variability of a large number of wind turbines aggregated together is fairly small at the regulation timescale. Additionally, load and wind power are uncorrelated on the small timescales considered for regulation and thus their fluctuations are rarely amplified and often cancel each other out. However, studies are continuing to use different methods to determine the impact to regulation reserve, and the requirements differ sometimes substantially.

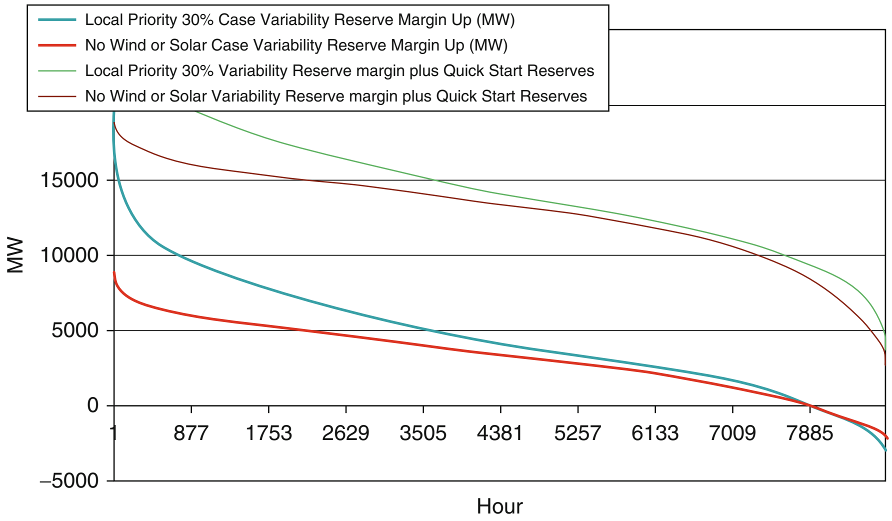
Load Following

One of the areas where the integration of renewable generation sources can have a major impact on grid operations is in the load following domain. In the minutes to hours time frame in which load following operates, there can be large ramps in renewable generation output. The addition of variable generation to load increases the overall system variability within this timeframe. For example, these effects occur when wind power output starts to ramp down following its diurnal pattern, just as the morning load starts to ramp up. This case marks extremely poor timing for the system operator as they must not only be able to handle the increase in load, but must simultaneously replace power that was being generated.

Wind Uncertainty Costs

The main costs of wind uncertainty are the result of having a suboptimal generation mix online due to an inaccurate forecast. One example of a commitment cost can be seen in the case of wind over-forecasting. If the wind output is forecast to be high, but is actually much lower in real time, a unit must be ready to serve the load that the wind was scheduled to serve. If this amount is significant it is possible that a slow starting baseload unit should have been made available at this time. In this case the wind overestimation causes an expensive fast starting unit to produce when a cheaper, but slow starting, unit would have been chosen if the forecast was more accurate. On the other hand, if the wind is under-forecast there may be more plants online than is necessary, causing higher start-up and fuel costs and leading to the possibility of wind curtailment.

It is important to note that since individual wind and solar plants are currently always smaller than the largest generator in the system, additional contingency reserves are not required by traditional metrics, such as the n-1 criteria. This could change in the future at higher renewable penetration rates if very large renewable generation plants are installed, or in the case of number plants being routed through a single large transmission line. However, at very high renewable



Renewable Generation, Integration of. Figure 6
Increase in available spinning reserves for a base case and a high renewable penetration rate case [5]

penetration rates it can be expected that very large renewable forecast errors could lead to situations where even contingency reserves are not sufficient to cover the error, as has been seen in a study of the western United States [5]. In the Western Wind and Solar Integration Study (WWSIS) it was found that these errors occur in approximately 1% of operating hours. While additional spinning reserves could be added to the system to handle these extreme events, it was proposed that some form of demand response would be more economic. While additional reserves of up to one to one backup have been proposed, this ignores the fact that the addition of renewable generation increases the amount of up-reserves available in the system. The incorporation of increased renewable sources causes the backing down of traditional generators, creating larger amounts of spinning reserve available to the system, as was shown in the WWSIS. One example of this result is shown in the up-reserve duration curve in Fig. 6. The amount of down spinning reserve is also increased as wind curtailment provides a very simple option.

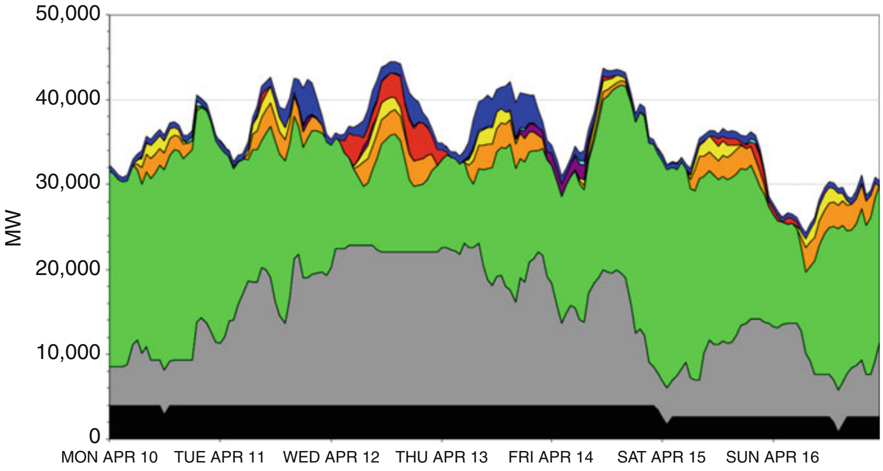
Thermal Unit Cycling

At very high penetrations of renewable generation it is not only peaking and intermediate units that are displaced.

Figure 7 shows an instance where baseload power is forced to produce at levels below maximum generation because of the large amounts of renewable power provided in a 30% renewable penetration scenario. When baseload generation must frequently change its level of output it is referred to as cycling, and it imposes additional wear and tear costs on units that were designed for near constant output. This particular instance is a case where a combination of low demand and high wind output combine to produce results that are far from normal system operation. The most damaging cycling events for thermal units are those where large changes in temperature cause material fatigue, decreasing the normal life of generator components. For this reason unit start-ups are the most damaging due to the large temperature differences between the operating mode and the shutdown mode. The length of the shutdown period is also important with hot-starts and warm-starts being preferred to cold-starts. Though less damaging than shutting down, bringing the generator down to its minimum output level, and then back up to maximum output, also creates large temperature changes that increase wear and tear on the unit.

Market Considerations

While it has been shown that low levels of renewable generation can be incorporated into the electricity



Renewable Generation, Integration of. Figure 7
An example of unit commitment and dispatch in the western United States with 30% renewable generation penetration [5]

system, with few or no changes to current system operations, the variable and uncertain nature of wind and solar power do create limitations on the penetration rates that are reasonably achievable within the context of the existing system. The effects of the uncertainty and variability are mitigated by large day-ahead and hour-ahead markets that provide many options for the balancing of supply and demand.

Balancing Area Cooperation

In 2010, there were over 100 balancing areas in the United States and Canada. Recent trends favor balancing area consolidation and the number of balancing areas is expected to continue to decrease. This is a positive development for the integration of renewable generation. Small balancing areas suffer larger consequences from routine variability. This is true for both supply and demand. For example, in a system with 100 MW total demand, an increase in demand of 10 MW is very significant, while in a system of 10,000 MW the 10 MW change is much more easily accommodated. Aggregated demand is also less variable for larger systems due to the larger number of individuals, and hence the smaller roles of the individuals in the aggregated system load. The variability of renewable generation also decreases with the aggregation of multiple generators in different locations. This geographical diversity of variable generation is due to the fact that the further the two generators are apart, the less likely they are to be affected by the same weather patterns, and hence the correlation of their power output is lower.

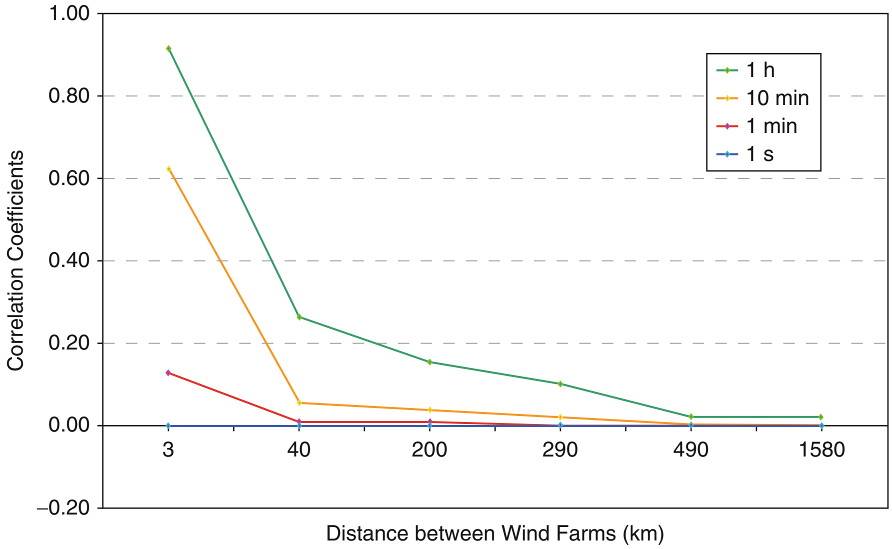
Figure 8 shows how the wind speed correlations decrease between geographically distant locations. Larger balancing areas will increase the potential area from which renewable generation can be drawn, and therefore lower the overall variability of total renewable output.

Reserve Sharing

Balancing areas may also choose to cooperate through other mechanisms besides full consolidation. One way that they may reduce the costs of variable generation integration is through reserve sharing agreements. Under these conditions the balancing areas independently schedule and dispatch generation to meet their own loads. However, since certain types of reserves, such as contingency reserves, are utilized only infrequently they may serve as backup to multiple balancing areas, should sufficient transmission capacity be available.

Dispatch Intervals

Another means by which markets may reduce the effects of system variability is through the use of more



Renewable Generation, Integration of. Figure 8
Wind speed correlations with distances between locations in the United States [14]

frequent market operations than the typical 1-h period for dispatch. Sub-hourly balancing markets aid in the integration of renewable generators by enabling the utilization of existing flexibility in the system and reducing the requirements for ancillary services. In addition, they allow the renewable generators to revise their production forecasts closer to the operating period to minimize forecast errors and thus imbalances.

Ancillary Service Markets

Ancillary service markets will play a key role in the integration of high levels of renewable energy resources. However, the current services offered may not be sufficient and new services should be created based on power system reliability requirements. For example, load following is an important system function with large amounts of renewable generators due to the increased system variability. While this is currently supplied through other units in the energy market, at high renewable penetration rates a market for fast ramping resources may prove to be advantageous. Another key consideration is the adoption of varying ancillary service requirements in response to current renewable generation conditions. When the amount of wind power currently being produced is high, the securing of large amounts of load following reserves would be wise, should a significant decrease in wind power output occur. However, when the wind power output is at a low level, the need for load following reserves is significantly decreased. By co-optimizing not only the choice of ancillary service providers with energy providers, but also the amount of ancillary services, the cost of variable generation integration can be reduced.

Capacity Markets

Well functioning electricity markets are required not only to supply the necessary amount of energy to meet demand at each point in time, but also to supply the generating capacity necessary. The calculation of the capacity credit for variable generation is one of the most contentious issues related to renewable integration, with many different methods being utilized in different systems [15]. Since a capacity credit is normally based on a system reliability measure, such as loss of load probability, it is effectively a measure of a units contribution to system reliability. Since the need for reserve capacity is highest at times of high system stress, when a loss of load is most likely to occur, capacity credit measures value available capacity at these times as more valuable than at times with a low probability of unserved load. Variable generators have a lower capacity credit than all but the most unreliable of conventional generators due to their high effective forced outage rate. These high effective forced outage rates are the result of weather-induced unavailability and not generator failures. Since the correlation between wind power output and peak load is weak in most locations wind generators tend to have fairly low capacity credit, as the timing of the capacity availability is so critical in the credit determination. On the other hand, solar generators tend to have relatively higher capacity credits, despite their generally lower overall capacity factors. This is due to the fact that the most regular instances of high solar output correspond to the same times as peak demand.

Locational Marginal Prices

Variable generators generally have high capital costs, but low variable costs when compared with conventional generators. The practical implication is that variable generation units will have lower marginal costs of production (often bid as zero) that they may submit as bids into energy markets. When the amount of variable generation is sufficient to remove the highest bidding unit that would otherwise be producing, the market cost of energy at that point in time is reduced. This reduces the total revenue collected by all generators and thus total system costs. Transmission system congestion leads to different prices at different nodes in the system, known as locational marginal pricing (LMP). When government policies for renewable energy production, such as production tax credits, are combined with very high renewable output at a particular point in the transmission system, the possibility of negative LMPs arises. In this case the production incentives offered make it economically efficient for units to pay to produce electricity. This situation is not sustainable in the long-term, neither for the variable generator nor for conventional units in the same area. If this situation occurs at a high enough frequency it could lead to the decommissioning of conventional units in the area, to the detriment of system reliability that relies on those units for reserve capacity [16].

Transmission Planning

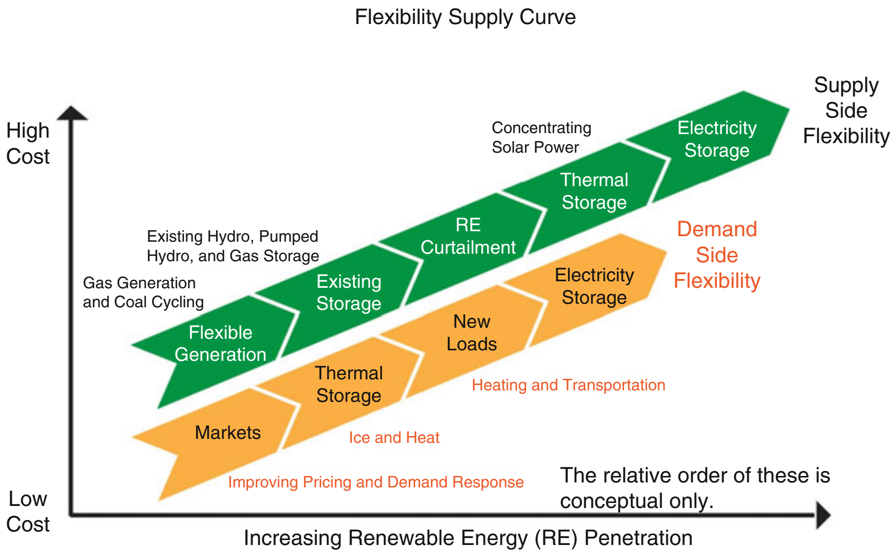
The sites that offer renewable generation sources the highest capacity factors are often located far from load centers. This dictates that new transmission must be built in order to utilize these areas with high renewable resources. While the need for new transmission is easily identified the process times for permitting and construction of the transmission often exceed those for the siting and construction of renewable generation. This can lead to situations where large amounts of renewable generation cannot be utilized because of transmission bottlenecks.

Transmission planning is not only important for gaining access to renewable resources, it can also be important in helping to integrate large amounts of renewable generation into the electricity system. Additional transmission between balancing areas or interconnects can allow for better utilization of renewable generation by providing alternative outlets for generation above what can be utilized within the local area. Having strong transmission ties between areas is a critical element in taking advantage of the geographic diversity of renewable resources. When the current renewable generation is low in one area but high in another additional transmission allows this energy to be utilized instead of curtailed. One proposed project in the United States that plans on utilizing increased transmission to aid renewable integration is the Tres Amigas project [17]. The North American electricity grid is split into three essentially independent interconnections: the Western, the Eastern, and Texas. The Tres Amigas project aims to provide expanded transmission links through the addition of 5 GW DC superconducting lines between each of the interconnections. The proposed site on the border of New Mexico and Texas is located close to areas of excellent wind and solar resources.

An even more ambition project for the integration of vast renewable resources over a large geographic footprint is the DESERTEC project [18]. This project aims to create a supergrid throughout Europe, the Middle East, and North Africa. This supergrid would open access to a larger number of renewable resources, including concentrating solar electricity from the Middle East and North Africa and wind power from Northern Europe. The enormous footprint of the project would ensure a smoothing of the renewable generation sources and demand sinks.

Enabling Greater Renewable Penetration

Just as there are currently a number of generation technologies utilized to fulfill electricity supply, there


Flexibility Supply Curve
High Cost
Low Cost
Supply Side Flexibility
Concentrating Solar Power
Electricity Storage
Thermal Storage
RE Curtailment
Existing Hydro, Pumped Hydro, and Gas Storage
Existing Storage
Gas Generation and Coal Cycling
Flexible Generation
Demand Side Flexibility
Electricity Storage
New Loads
Heating and Transportation
Thermal Storage
Ice and Heat
Markets
Improving Pricing and Demand Response
The relative order of these is conceptual only.
Increasing Renewable Energy (RE) Penetration

Renewable Generation, Integration of. Figure 9
The flexibility supply curve shows a number of different methods of providing system flexibility that may be useful in integrating renewable generation sources [19]

will be a number of technologies and strategies needed in order to enable the inclusion of larger amounts of renewable generation into the electricity system. Figure 9 shows a conceptual cost ranking of different demand and supply side technologies and systems that provide the electricity system with additional flexibility, and thus can help in the integration of variable generation. The inclusion of a number of different renewable generation technologies, with complementary generation characteristics, in future capacity expansion will also help to facilitate operating a high renewable electricity system. Dispatchable renewable sources, such as geothermal, biomass-fired thermal and hydroelectric plants, are able to contribute to renewable generation goals without significantly altering system operations. Geothermal and biomass-fired plants can serve as baseload units that reduce the amount of variable generation necessary at all times. Hydroelectric units can also serve this function, but may better serve the system by providing load following capabilities for variable generation due to their quick response times. However, the limited geographic potential and/or economic costs of these plants dictate that large capacities from variable sources will also be required to meet high renewable penetration goals. A number of different strategies for mitigating the effects of these plants' variability and uncertainty are described in what follows.

Variable Generation Forecasting

Renewable generation is both variable and uncertain; however, the uncertainty is a much more critical factor than the variability. If the variable output of renewable generation was known in advance the impact would be considerably reduced. The variable output could still be scheduled in a normal unit commitment and dispatch system, and would in fact be dispatchable. It is the uncertainty associated with renewable output that is most troubling. Thankfully, the future output of a variable generator is not completely random at smaller timescales, due to the fact that it is weather-driven. Forecasting techniques can be used to estimate the future output and this information can be incorporated into future generation plans. However, forecasting results tend to improve with reduced lead times between making the forecast and the time of realization. The inaccuracies associated with forecasting at longer timescales, in conjunction with the long start-up times of baseload units, diminishes the practical impacts of wind forecasting in particular. Numerical weather prediction models are commonly used to make wind forecasts for the day-ahead unit commitment process. Further improvements to these models, and their use at smaller geographic distances, has the potential to improve the utilization of variable generation through better forecasting.

Stochastic Planning and Operating Tools

One way in which variable generation forecasting can be better utilized is through the use of stochastic unit commitment and economic dispatch models. Instead of a simple point forecast these models incorporate a number of different scenarios of variable generation power output, with associated probabilities, during the future time frames under consideration. Through their explicit consideration of the stochastic nature of the variable generation these models are able to produce more robust schedules that can respond more easily to the different possible variable generation scenarios. The more robust schedules produced tend to produce lower average system costs than those that utilize only a point forecast [20]. Improved forecasting that considers not only the most likely value for the next time point, but also a consideration of the range and likelihood of possible values will increase the impact of stochastic operating tools. For this reason the consideration of the distribution of forecasting errors for variable generation production is important as it can inform the creation of more realistic forecasting scenarios. Improved stochastic programming algorithms that reduce the computational time necessary to solve for an optimal schedule will also increase the effect of these tools by allowing for the consideration of a larger number of scenarios and their application at smaller timescales.

Faster Markets

As the time between the forecast of variable generation and the actual production time decreases, so does the average error of the forecast. For this reason the implementation of faster dispatch markets can help lead to further renewable generation penetration as it decreases the costs associated with variable generation uncertainty. By having less time between the time of forecast and the actual time of output, more current information can be utilized in deciding the forecast. Since day-ahead markets followed by 1-h dispatch market tend to be the normal historical system of market operation, sub-hourly markets are necessary for further improvements. These sub-hourly balancing markets also provide another means by which the regulation and load following reserve impacts of variable generation may be reduced.

Demand Response

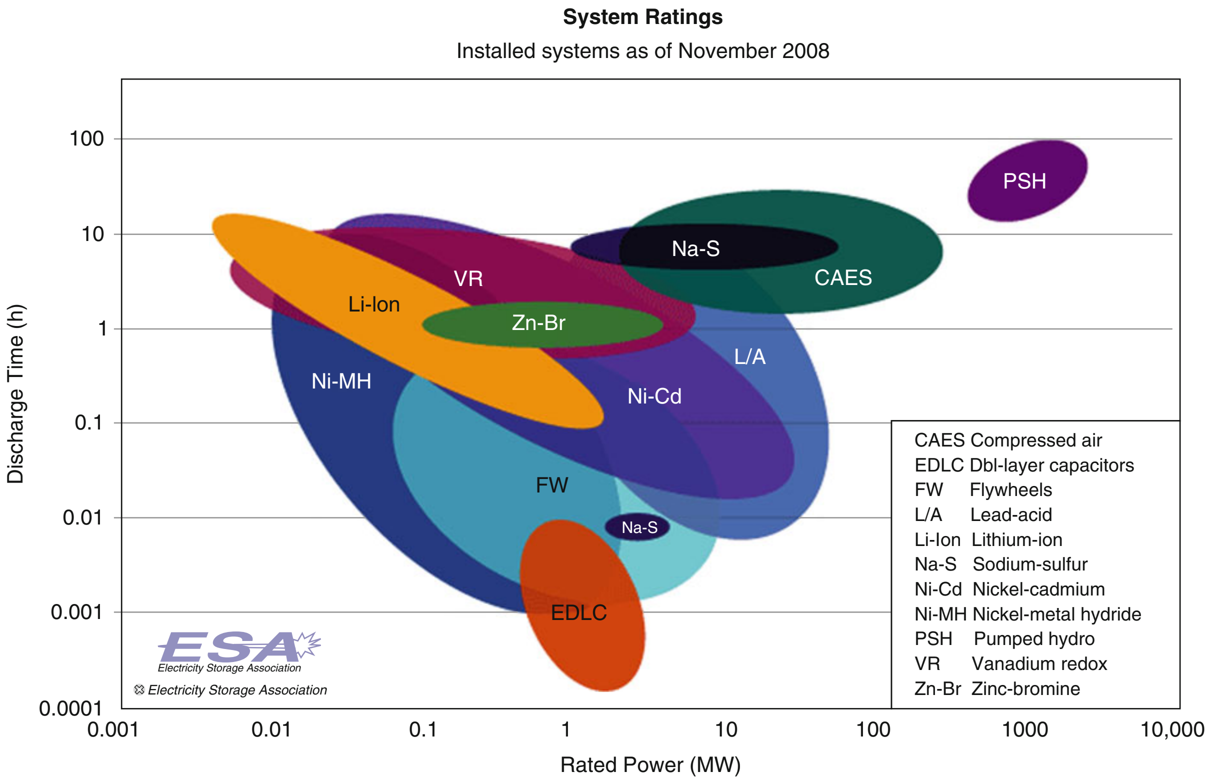
Another possible approach to incorporating larger fractions of renewable generation into the electricity system is known as demand response. This approach is based upon changing one of the main tenets of traditional electricity system operation: varying generation to match an uncontrollable load. Demand response gives the system operator control over the timing of some portion of load by allowing blocks of load to be delayed. This approach has traditionally been used by system operators as a last resource during times of very high peak loads. Large industrial loads often structure their contracts with utilities so that they receive either payments or lower base rates for allowing the utility the privilege of interrupting their service periodically, up to a maximum number of occurrences. Industrial customers are preferable from the utilities' perspective because they allow a large reduction in load through a single customer interruption. Large commercial installations also have the possibility of participating in demand response programs, as the heating or cooling of a large commercial building could be a significant resource.

The adoption of "SmartGrid" technologies has the promise of allowing residential customers to also participate in demand response programs. Unlike industrial loads, individual residential customers are a very small percentage of the total system load. However, if the utility can simultaneously control a large number of large-load distributed residential appliances, such as air conditioners, the aggregated effect can be comparable to the response provided by a large load industrial customer. It is important to note that demand response programs do not significantly reduce total electricity usage, instead the load is shifted from times of high demand, when additional generation may be unavailable, to periods of lower demand. Household demand response is limited by cost to large appliances, and by consumer acceptance to appliances whose usage patterns are somewhat time insensitive. For example, a refrigerator's load can be postponed, but only for very short timescales, so that the contents do not spoil. On the other hand, a dishwasher turning on can presumably be delayed until the total demand levels fall at night, with far less negative effects to the consumer. In the extreme case one can imagine the linking of certain appliances' usage to the current level of variable generation output. A related idea is the use of time of use pricing for residential electricity, as opposed to the flat-rate (or limited peak/off-peak split) rate structure that currently prevails. The most troublesome aspect of time of use pricing is the limited ability or desire of residential customers to react to different prices. For extensive changes in residential usage patterns the time of use pricing scheme would need to be paired with meters that can be programmed to adjust appliance usage patterns based on the current price of electricity.

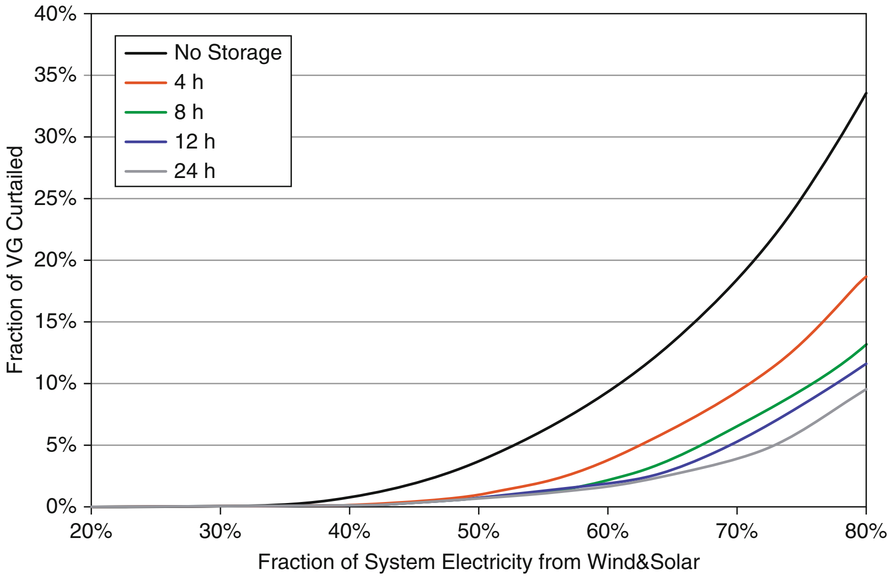
Electricity Storage

A commonly proposed solution to mitigate the effects of renewable generation variability is the use of electricity storage. There are a number of different technologies that could be used for such purposes, with each technology operating most efficiently over a range of timescales and power and energy levels. An illustration of the effective ranges for different technologies can be seen in Fig. 10. There are three basic timescales on which electricity storage is needed, based on response duration: very short duration, short duration, and long duration. Generally speaking, the longer the timeframe is, the higher the associated energy requirements are. These types of storage can also be classified according to their application, where power quality, bridging power, and energy management correspond well to very short, short, and long duration timeframes, respectively [21]. Very short duration storage is needed to respond on the millisecond timescale and provide large amounts of power for a matter of seconds. This type of power is important for power quality and frequency regulation applications. Requiring a response on the multiple second to minute timescale is short duration storage, used for bridging power. This type of storage is useful in the role of generation reserve, and similar applications, to provide power in the event of a system contingency. Very long duration storage can respond on the multiple minute timescale and is used to provide power over long durations, such as those needed for load leveling applications.

While both bridging power and power quality applications are important considerations in integrating variable generation, storage technologies with large rated power output and long discharge times will be the most important for large penetration rates of renewable generation due to their ability to shift energy from times when variable output exceeds demand to those where demand is greater than variable output. Pumped hydroelectricity is one large-scale technology that is applicable for long timescales. During times of excess generation the pumped hydro facility can transfer water from a lower reservoir to a higher reservoir and utilize the potential energy gained during times of high demand. Pumped hydroelectricity storage is somewhat limited by geographic considerations; there must be a fairly large height difference between the two reservoirs, ruling out very flat locations. Compressed air electricity storage is also limited by geographic considerations, particularly for systems with long discharge times. This is due to the fact that very large systems rely on caverns as the compressed air storage vessel. In these systems excess power is used to compress air into a storage area, where it can later be expanded to produce electricity when desired. Current systems use natural gas combustion to supplement the storage compressed air, but designs for future systems need to include this option. Chemical storage in the form of batteries is another possible solution. Lead acid batteries are currently a cost-effective solution, but are also a very mature technology that have relatively low roundtrip efficiency and are limited in grid applications due to their relatively short lifetimes. Sodium-sulfur (NaS) batteries have long cycle life, relatively high efficiency, and are used primarily for grid applications due to the high operating temperatures. A number of other battery chemistries are also available, from those such as lithium-ion that are seen in many other applications where portability is essential, to technologies such as redox flow batteries that are exclusive to large-scale power applications.



Renewable Generation, Integration of. Figure 10
The range of power and energy combinations available for different storage technologies



Renewable Generation, Integration of. Figure 11
Total curtailment as a function of VG energy penetration for different amounts of energy storage. Assumes 30/70 solar/wind mix, 12 h of storage, and a 100% flexible system. Each hour of storage represents 1 h of average system demand [19]

Figure 11 shows the effect that electricity storage can have on efficient system operation, in terms of reduced wind curtailment, in a theoretical completely flexible system. As may be seen, even a relatively small amount of storage can greatly increase the potential renewable penetration rate by reducing the amount of variable generation that must be curtailed, thus improving the economic viability of additional variable generation capacity. Additional amounts of energy storage can aid in the integration of renewable generation, but provide diminishing marginal returns. Currently electricity storage options are fairly expensive, limiting their potential applications. Even if significant cost improvements in storage technologies are realized, electricity storage should be viewed as but one tool to aid in renewable generation integration.

Renewable Generation Curtailment

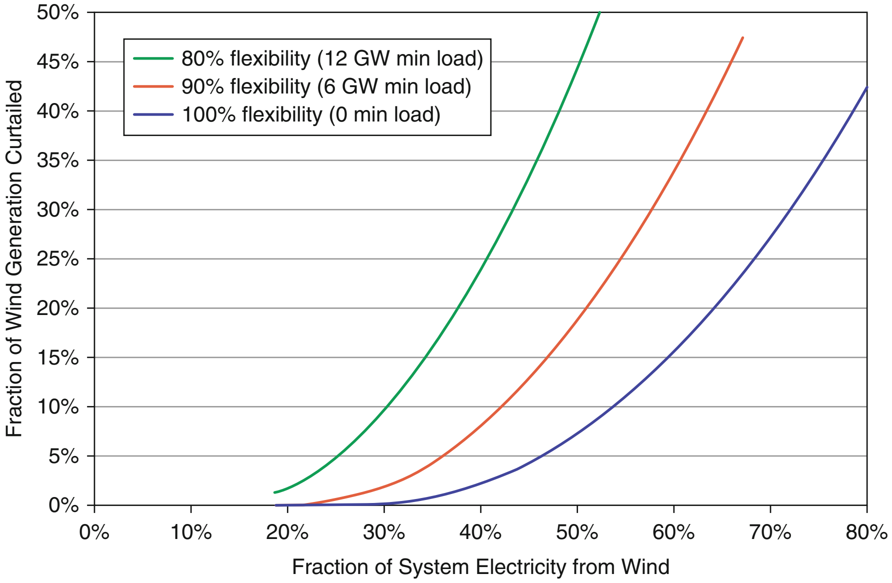
Another solution that is occasionally used, even at the current low levels of renewable generation penetration, is production curtailment or spilling. If the production from variable generation is so large that it affects the operation of other units, such as putting baseload production below minimum generation levels, or reaches physical constraints, such as transmission line capacities, then the production can be temporarily halted. While this is a simple solution for small amounts of variable generation, it has its limits of applicability. These are often economic, in that excess variable capacity can be built that will only produce energy at a very small number of needed time points, and will be curtailed during the rest of its possible production times. This would significantly affect the economics of the variable generation plant at very high levels of curtailment, making it economically inefficient to build further capacity unless the load production schedule is expected to be anti-correlated with other variable generation locations.

New Loads

The peak and trough cycle of daily electricity usage requires the use of different types of generation in order to maintain system flexibility. However, the system would be able to operate more efficiently if the daily usage profile was more flat and baseload generators could provide a larger fraction of total energy. This would require sources of new load whose usage patterns coincided with the current nightly trough. In the same vein, since wind power output is generally higher at night, new nighttime loads would allow for less wind curtailment as there would be fewer situations where wind plus minimum baseload generation exceeded the current demand. The most commonly named source of a new load that can serve these purposes is the electric vehicle. Most vehicle usage occurs during the day, leaving vehicles idle during the night. For electric vehicles this provides the opportunity to charge and have their full range capabilities in time for the owner's use in the morning. The charging pattern that the vehicles follow is very important in ensuring that they flatten the load profile, instead of increasing peak loads [22]. To this end, policies must be put in place that can benefit the electricity system without disturbing the benefits that consumers derive from the vehicles, and thus blunting the rate of adoption. One extreme example of a policy intended to derive the most system benefits from the vehicles is the idea of utility controlled charging. In this case the utility would be able to use the vehicles as a form of demand response, perhaps providing an outlet for variable generation that might otherwise be curtailed.

Flexible Generation

The current electricity generation portfolio was not designed with the incorporation of variable generation resources in mind. Very high levels of variable generation penetration will most likely require more flexible accompanying generation, in order to help compensate for the non-dispatchable output. Generator flexibility includes the ability to both start and ramp quickly. For example, current inflexible nuclear and coal plants can require hours to ramp up from a cold start to full capacity. On the other hand, simple cycle natural gas turbines are an example of a unit that can start quickly and ramp from minimum generation levels to full capacity quickly enough to be used to compensate for generator outages. While these flexible units provide the system operator with more dynamic options for meeting the load, they also tend to be more expensive than inflexible baseload plants. However, the current level of system flexibility is not fixed. As older inflexible units reach the end of their operating lives and are



Renewable Generation, Integration of. Figure 12
Total curtailment as a function of usable wind energy penetration for different system flexibilities [19]

retired, new more flexible units can be brought online to replace them. The impact of system flexibility on the integration of variable generation is shown in Fig. 12. Here flexibility is represented solely by the combined minimum load levels of all generation units in the system. The ability to integrate variable generation is represented by the amount of available wind generation that must be curtailed as a percentage of the fraction of the total system energy provided by wind power. As may be observed, increased system flexibility causes the amount of wind generation curtailment to drop significantly.

Future Directions

Variable generation penetration rates are still relatively low (below 10% by energy) in most large systems. At these lower levels variable generation can be fairly easily incorporated into existing electricity system operations. However, if global trends persist very significant penetration rates will soon be reached. At high penetrations of variable generation significant restructuring of current system operations could be necessary to accommodate the additional system variability and uncertainty. As many electricity systems reach penetration rates of between 15% and 30%, methods of economically and reliably integrating variable generation sources will start to be tested on a daily basis. This will quickly bring attention to issues that were not properly considered in previous integration studies. Potential operating issues at higher levels of variable generation must be anticipated before these higher levels of renewable penetration are realized in order to maintain system reliability. The electricity system is such a vital part of daily life that even small changes in the reliability of the system would have far-reaching economic and societal consequences. Further work is required in many areas of renewable generation integration. Resource assessment of wind and solar resources can lead to better decisions on where to site new generation capacity to maximize not only power output, but the benefit to the system. Technological development of generators can both reduce system costs and allow for variable generators to interact more smoothly with the traditional system operation paradigm. Further work on the characterization of variable generators is necessary to understand how systems with large penetration rates will behave. Finally, extensive work is necessary in the area of system operations. Understanding the ability of both supply and demand technologies to increase system flexibility will be critical, as will new ideas on the structure of systems with large amounts of variable generation.

Bibliography

Primary Literature

1. NERC (2009) Accommodating high levels of variable generation. North American Electric Reliability Corporation. The North American Electric Reliability Corporation is the publisher. The report may be found at: http://www.nerc.com/files/IVGTF_Report_041609.pdf
2. IEA (2010) World energy outlook 2010. International Energy Agency, Paris
3. EIA (2011) Electric power annual 2009. Energy Information Administration, Washington, DC
4. Denholm P, Ela E, Kirby B, Milligan M (2010) The role of energy storage with renewable electricity generation. National Renewable Energy Laboratory, Golden
5. GE (2010) Western wind and solar integration study. National Renewable Energy Laboratory, Golden, 536pp
6. Milligan M et al (2010) Operating reserves and wind power integration: an international comparison. The 9th Annual International Workshop on Large-Scale Integration of Wind Power into Power Systems, Quebec, Canada
7. Milligan M, Kirby B (2009) Calculating wind integration costs: separating wind energy value from integration cost impacts. National Renewable Energy Laboratory, Golden
8. Lu X, McElroy M, Kiviluoma J (2009) Global potential for wind-generated electricity. Proc Natl Acad Sci USA 106:10933–10938
9. Schwartz M, Heimiller D, Haymes S, Musial W (2010) Assessment of offshore wind energy resources for the United States. National Renewable Energy Laboratory, Golden
10. de Vries B, van Vuren D, Hoogwijk M (2007) Renewable energy sources: their global potential for the first-half of the 21st century at a global level: an integrated approach. Energ Policy 35:2590–2610
11. Sims R et al (2007) Energy supply. In: Metz B, Davidson O, Bosch P, Dave R, Meyer L (eds) Contribution of working group III to the fourth assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge
12. Chen Z (2005) Issues of connecting wind farms into power systems. In: IEEE/PES transmission and distribution conference and exhibition: Asia and Pacific, Dalian
13. GE (2005) The effects of integrating wind power on transmission system planning, reliability, and operations: phase 2: system performance evaluation. The New York State Energy Research and Development Authority, Albany
14. Wan Y-H (2005) A primer on wind power for utility applications. National Renewable Energy Laboratory, Golden
15. Milligan M, Porter K (2006) The capacity value of wind in the United States: methods and implementation. Electr J 19:91–99
16. Smith JC et al (2010) Impact of variable renewable energy on US electricity markets. In: IEEE power and energy society general meeting, Minneapolis
17. Z-Global (2010) Study of the economic impact of Tres Amigas. Western Electricity Coordinating Council – Transmission Expansion Planning Committee. The Western Electricity Coordinating Council is the publisher. The report may be found at: http://www.wecc.biz/committees/BOD/TEPPC/TAS/MWG/Shared%20Documents/Tres%20Amigas%20Modeling/Z-Global%20Analysis.pdf
18. DESERTEC (2007) Clean power from deserts – the DESERTEC concept for energy, water and climate security. Protext, Bonn
19. Denholm P, Hand M (2011) Grid flexibility and storage required to achieve very high penetration of variable renewable electricity. Energy Policy 39(3):1817–1830
20. Tuohy A, Meibom P, Denny E, O'Malley M (2009) Unit commitment for systems with significant wind penetration. IEEE Trans Power Syst 24(2):592–601
21. Nourai A (2002) Large-scale electricity storage technologies for energy management. In: IEEE power engineering society summer meeting, Chicago
22. Hodge B-M, Shukla A, Huang S, Reklaitis G, Venkatasubramanian V, Pekny J (2011) A multi-paradigm modeling simulation of the effects of PHEV adoption on electric utility usage levels and emissions. Ind Eng Chem Res 50(9):5191–5203

Books and Review

Ackermann T (ed) (2005) Wind power in power systems. Wiley, Chichester

Anaya-Lara O, Jenkins N, Ekanayake J, Cartwright P, Hughes M (2009) Wind energy generation: modeling and control. Wiley, Chichester

Blume S (2007) Electric power system basics: for the nonelectrical professional. IEEE/Wiley, Hoboken

Brinkman G, Denholm P, Drury E, Margolis R, Mowers M (2011) Toward a solar-powered grid: operational impacts of solar electricity generation. IEEE Power Energy Mag 9–3:24–32

Burton T, Sharpe D, Jenkins N, Bossanyi E (2001) Wind energy handbook. Wiley, Chichester

Carrasco JM, Garcia Franquelo L, Bialasiewicz J, Galvan E, Portillo Guisado RC, Martin Prats MA, Leo JI, Moreno-Alfonso N (2006) Power-electronic systems for the grid integration of renewable energy sources: a survey. IEEE Trans Ind Electron 53:1002–1016

Corbus D, Lew D, Jordan G, Winters W, Van Hull F, Manobianco J, Zavadil B (2009) Up with wind: studying the integration and transmission of higher levels of wind power. IEEE Power Energy Mag 7–6:36–46

Freris L, Infield D (2008) Renewable energy in power systems. Wiley, Chichester

Grant W, Edelson D, Dumas J, Zack J, Ahlstrom M, Kehler J, Storck P, Lerner J, Parks K, Finley C (2009) Change in the air: operational challenges in wind-power production and prediction. IEEE Power Energy Mag 7–6:47–58

Grubb MJ (1991) The integration of renewable electricity sources. Energ Policy 19:670–688 von Meier A (2006) Electric power systems: a conceptual introduction. IEEE/Wiley, Hoboken

Miller R, Malinowski J (1993) Power system operation. McGraw Hill, Boston

Mills A, Ahlstrom M, Brower M, Ellis A, George R, Hoff T, Kroposki B, Leox C, Miller N, Milligan M, Stein J, Wan Y (2011) Dark shadows: understanding variability and uncertainty of photovoltaics for integration with the electric power system. IEEE Power Energy Mag 9–3:33–41

Shahidehpour M, Yamin H, Li Z (2002) Market operations in electric power systems. IEEE/Wiley, New York

Smith JC, Milligan MR, DeMeo EA, Parsons B (2007) Utility wind integration and operating impact state of the art. IEEE Trans Power Syst 22:900–908

Soder L, Hofmann L, Orths A, Holttinen H, Wan Y, Tuohy A (2007) Experience from wind integration in some high penetration areas. IEEE Trans Energy Convers 22:4–12

Reservoir Engineering in Geothermal Fields

Enrique Lima, Hiroyuki Tokita, Hideki Hatanaka
West Japan Engineering Consultants, Inc, Chuou-ku,
Fukuoka, Japan

Article Outline

Glossary
Definition of the Subject and Its Importance
Introduction
Purposes of Reservoir Engineering in Geothermal Fields
Information and Processing Required for Geothermal Reservoir Evaluation
Numerical Simulation
Resource Evaluation
Future Directions
Bibliography

Glossary

Aquifer A geological formation (or formations) which contains water or geothermal fluid and will allow fluid movement.

Baseline Data set acquired before exploitation begins, against which any future measurements are compared.

Deep liquid zone Region of single-phase liquid conditions below a two-phase (liquid and vapor) zone.

Deep liquid level Boundary between the two-phase and deep liquid zones.

Energy Conversion technology Term that refers to the thermodynamic cycle used to convert the heat energy from geothermal fluids into electricity, there are several of these technologies. Binary in which the heat of the geothermal fluids is transferred to a low-vaporization-temperature fluid (usually isopentane, ammonia, or a combination of water–ammonia) that is the working fluid driving the turbine generator. The heat source for the binary technology could be the separated geothermal brine or the steam or both exchanging energy in separated exchangers. When the steam is used directly, it can be used in single-flash or double-flash cycles, depending on the number of times the geothermal brine is flashed into steam or a combination of binary and steam technologies in the conversion cycle.

Geothermal system A body of hot water and rock within the Earth.

Groundwater Water, generally cold and of meteoric origin, which resides in near-surface aquifers and is often used for domestic and industrial purposes.

Injection zone The formation into which injected fluid is put. Generally, this has high porosity and permeability.

High-temperature system A geothermal system, or part thereof, containing fluid having a temperature greater than 150°C. c.f. *Low-temperature system* in which the temperature is less than 150°C. Note, however, that this temperature value is arbitrary and that different authorities adopt different values or divide the range into low, intermediate, and high temperature.

Makeup well Well drilled to replace production lost from an existing production well due to decreases in fluid temperature or pressure or permeability of the production zone, or drilled to replace reinjection lost from an existing reinjection well because of a decrease of permeability in the injection zone(s) or excessive rise of water level within the well.

Model In the reservoir engineering terminology, there are two kinds of models: one is called the *conceptual* *model* that is an integration of all the physical features of a reservoir such as size, geological structure, geochemical features, temperature and pressure, underground fluid flow patterns, etc., that is estimated through the integration of all available surface and underground geoscientific information. The other kind of model is called the *numerical model*, and it is the conceptual model itself presented in terms of elements that can be manipulated using numerical tools in order to extrapolate, from known information and following strict physical–mathematical rules, features of the underground that cannot be directly measured, such as volume, recharge and discharge zones, pressure and temperature distribution, energy content, etc., with the objective to mathematically forecast the response of the reservoir when subjected to different exploitation scenarios. Numerical models are constructed in three dimensions (x, y, and z) this is the common denomination of 3D numerical model. These numerical exercises of reservoir modeling will lead to crucial decisions such as to develop or not to develop the geothermal field and if decision to develop is made, then what would be the size of facilities that, in terms of economy, will produce an acceptable return to the country and investors.

Permeability A measure of the capacity of a geological rock formation to transmit a fluid.

Production zone That region (depth) of the geothermal reservoir from which most of the production of fluid occurs.

PT Pressure and temperature wellbore surveys could be static (not flowing conditions) or dynamic (flowing conditions). When a spinner tool is added in the dynamic surveys, the activity is called PTS wellbore survey.

Reservoir The region of a geothermal system from which geothermal fluid is withdrawn or is capable of being withdrawn.

Residual (liquid) saturation The amount of liquid that remains in the pores (as % of pore volume) and which decreases in pressure will not vaporize. The liquid saturation level below which vaporization of liquid will not occur.

Steam zone A region of the reservoir in which steam (vapor) is the pressure-controlling phase.

Two (2)-phase zone A region where the liquid and vapor (steam) phases of water coexist in pores or fractures.

Definition of the Subject and Its Importance

Reservoir engineering is the comprehensive integration of all available surface and underground information regarding the geology, geophysics, geochemistry, well drilling-testing, exploitation data, and information concerning the developer objectives, such as market targets, costs, and finance becoming the most power tool to evaluate the feasibility of a geothermal development. As in any scientific or engineering activity, results derived from reservoir engineering depend upon the quantity and quality of information and in the ability (experience) in handling all this information. Reservoir engineering is not limited to the final numerical tool but to the information acquisition, too. The present description encompasses both acquisition and processing of information to guarantee high-quality results.

Introduction

Among all forms of primary energies, renewable and nonrenewable geothermal energy is the only one that, from the perspective of a country together, offers values, such as energy source reliability, energy independence, contribution to improved commercial balance, clean energy, and local socioeconomic development due to multipurpose utilization possibilities. Again, from the point of view of energy mix, it is necessary to evaluate the potential available of the geothermal resource to plan where and when geothermal energy has to be deployed in order to produce the best effects of its inherent values. Planning should consider the risk and long lead time required to develop geothermal resources. Reducing lead time and eliminating or at least mitigating risks are among the aims of geothermal engineering. The development of geothermal energy is carried out in several steps until its utilization is possible: (1) national reconnaissance, (2) prefeasibility studies, (3) resource feasibility studies including exploration and appraisal drilling, (4) facilities < technical-economic/financial > feasibility studies, (5) financing acquisitions, (6) field development and facilities construction, and (7) facilities operation. Each of these steps is concatenated to the success of the preceding one, and each of them depends on the quantity and quality of the acquired data and its processing. The ultimate tool for the evaluation of the geothermal potential and the definition of the feasibility of its exploitation is reservoir engineering. The application of this tool is most relevant in steps 3 and 4 and is the basis for the success of step 5. The tool continues to be applied in step 6 and during all the exploitation period, step 7.

Purposes of Reservoir Engineering in Geothermal Fields

Reservoir engineering will be used for the following purposes:

1. Data acquisition
   (a) Wellbore pressure–temperature and mass flow surveys
   (b) Completion tests
   - Pressure transient tests
   - Injectivity tests

   (c) Well production test
   - Mass flow
   - Enthalpy
   - Chemical characteristics

   (d) Well interference test
   (e) Tracer tests
2. Data processing
   (a) Wellbore data analysis to obtain:
   - Static pressure–temperature profiles to determine stabilization temperature and reservoir pressure
   - Dynamic pressure–temperature profiles to determine production characteristics of the permeable zones
   - Position of permeable zones and evaluation of their thermal–hydraulic characteristics
   - Construction of the conceptual model
   - Calibration of the numerical model

   (b) Analysis of well completion data
   - Pressure transient data to determine permeability, storativity, and skin effect of formation around the permeable zones
   - Injection capacity of the well
   - Construction of the conceptual model
   - Data for the numerical model

   (c) Analysis of well production data
   - Construction of the production characteristics curve to determine production characteristics and chemistry at different well pressures and, in turn, to help to determine optimum well exploitation pressure and sizing of facilities
   - Calibration of the numerical model

   (d) Reservoir simulation
   - Estimation of the spatial pressure, temperature, and vapor saturation distribution within the reservoir volume
   - Estimation of the geothermal potential of the reservoir
   - Estimation of the strength and location of recharge, discharge, and heat source areas
   - Estimation of the location of promising permeable areas for production and reinjection
   - Estimation of the optimum sustainable exploitation capacity of the reservoir
   - Contribute to the optimal positioning of geothermal facilities (well pads and pipelines for both production and reinjection, and energy conversion facilities)
   - Contribution in the selection of energy conversion technology
   - Estimation of the optimum wellhead pressure for exploitation of the reservoir
   - Targeting and optimization of the number of initial production and reinjection wells
   - Targeting and optimization of the number of makeup production and reinjection wells
   - Estimation of development and construction costs
   - Contribution to the economic and financial study
3. Financing acquisition
   - Bankable report
4. Reservoir monitoring
   - Updating of reservoir analysis and forecasting of future performance as experience and new data are gained during exploitation of the field

Information and Processing Required for Geothermal Reservoir Evaluation

To assess the capability of reservoir fluids to produce electricity or to be used in other applications, it is necessary to gather and analyze information on the thermodynamics of the fluids, hydrogeological characteristics of the reservoir, and connectivity of the formation through which geothermal fluids are migrating. Thus, information on the temperature, pressure, enthalpy, and chemistry of the geothermal fluids and information on the location of permeable zones and permeability of the geological formations shall be collected and analyzed.

Temperature

Since the measure of the heat content in geothermal fluids is temperature, this parameter is of paramount importance in geothermal developments. The higher temperature fluid is, so will be the heat content. In terms of electricity generation, there is an appropriate range of temperatures of geothermal fluids at which energy conversion using a determined technology (from water-binary to double-flash cycles). At current commercial energy conversion technologies, the range of reservoir temperature in use is from around 90°C to over 300°C.

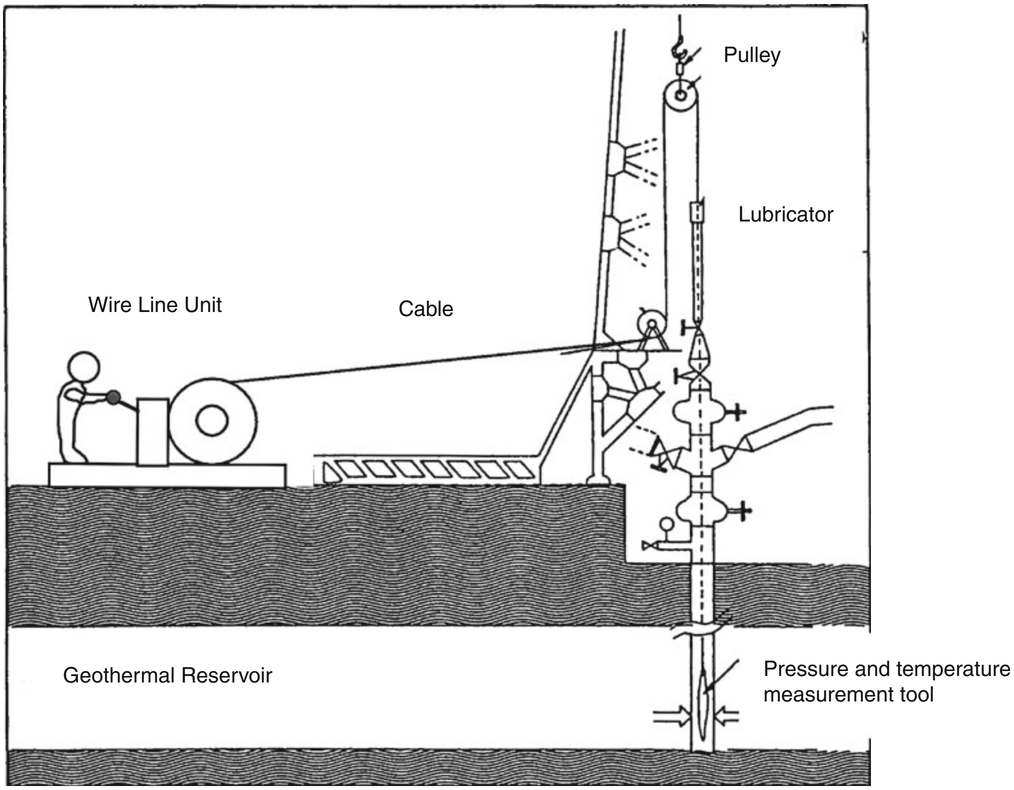
Reservoir temperature, therefore, becomes a parameter of importance to be measured. When deep wells are available, this is done by lowering instruments into the wellbore (Temperature Wellbore Survey), or when wells are not available, temperature is indirectly estimated using geochemical methods. Downhole temperature measurements are done at several times during and after drilling to estimate the thermal conditions around the surveyed well at undisturbed conditions (static surveys) and, most importantly, to determine the position of permeable zones along the open portion of the wellbore. The spatial distribution of the measurements in different wells is a guide to understanding the position of the heat source and flow patterns within the reservoir. When these measurements are done during drilling, the information serves to make important decisions regarding drilling operations. Nevertheless, suspension of drilling operations to collect downhole temperature represents costs such as the one to keep the rig in standby; thus, a compromise between the importance of data and cost should be attained. Other temperature measurements which are necessary to evaluate the reservoir are those done within the wellbore while the well is in production (dynamic surveys). These measurements help to understand the variations of enthalpy when the well is subject to mass production at different conditions of wellhead pressure. Temperature measurements are continuously done in the several wells of the development and during the life of the geothermal exploitation because the spatial control of temperature is an excellent measure of the form in which the reservoir is responding to exploitation.

Pressure

It is almost a standard that together with temperature, wellbore pressure measurements are also done. The combination of temperature and pressure profiles along the well (static and dynamic) provides information about the thermodynamic state of the geothermal fluid, which in turn renders information regarding the deep liquid zone (static), flashing point (dynamic), and flow patterns within the wellbore. Pressure information can be also utilized in combination with the discharge data for the further understanding of the flow mechanism and productivity of the well.

Pressure of geothermal reservoir is one of the factors that control the productivity of the well. The higher the reservoir pressure is, the shallower the static water level within the well will be. This translates into a higher productivity of the well because as the water level is at shallow depths, the distance of the reservoir fluids to the ground surface is shorter, thus requiring less energy for geothermal fluid to reach (ascend) to the ground surface.

Wellbore pressure and temperature are measured as shown in Fig. 1. Since the pressure is the cumulative weight of liquid inside the wellbore per unit area, the static pressure profiles are quasi-straight lines, as shown (Fig. 2). The slope of the pressure profile shows variations due to variations in fluid density caused by the temperature gradient within the wellbore. Static pressure profiles during the heating up of the well are useful to estimate the position of permeable zones. The pressure profile measured at long standing time that assures stabilization of the temperature profile is used to estimate reservoir pressure. The stabilized pressure and temperature profiles are in turn used to calibrate the 3D numerical model.

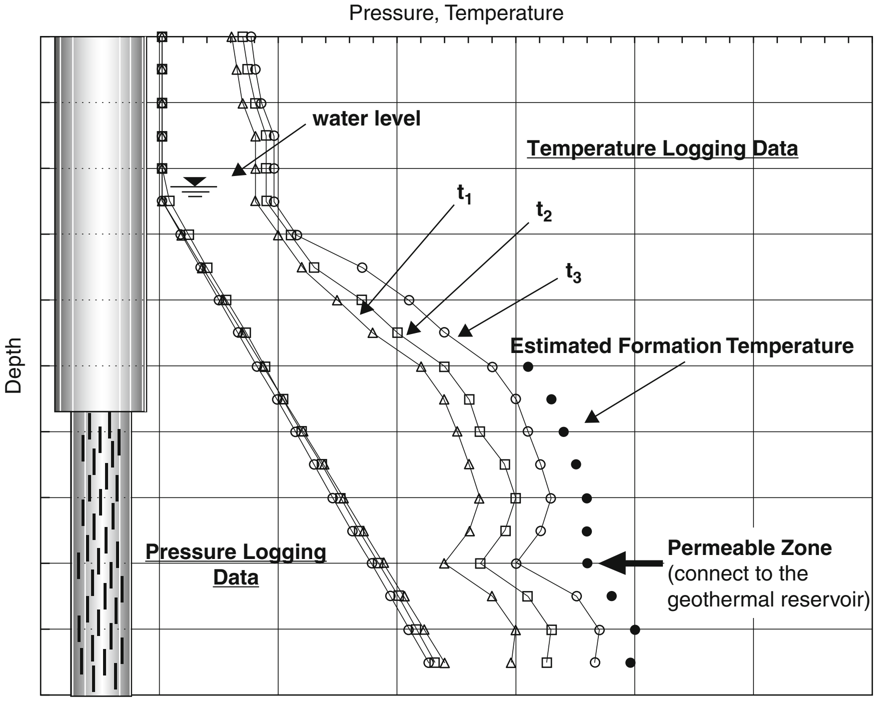


Reservoir Engineering in Geothermal Fields. Figure 1
Schematic of pressure and temperature logging

When pressure profiles of several wells are available, a spatial pressure distribution map of the field at different elevation levels can be created. Since the fluid basically moves from the high- to low-pressure environment, the pressure distribution map, together with the geological structure map and geochemistry information, can be used to estimate flow patterns of the fluid within the reservoir. Flow patterns information is of extreme importance during the calibration of the numerical model.

Wellbore pressure profiles in dynamic (flowing) conditions show lower values compared to those taken at static condition because of the frictional pressure drop within the wellbore and because of the appearance of a pressure drop between the reservoir and the production zone caused by the flow of the fluid within the geological formations. The degree decrease of pressure values is a good indicator of the "permeability" of the well. Variations in gradient of the pressure profile in dynamic conditions also are an indicator of the position of permeable zones and whether the fluids enter the wellbore in liquid, two-phase, or steam conditions. If the fluids enter the wellbore in liquid state, the position of the flashing zone can also be estimated from the dynamic pressure profile and calculated using information of the temperature profile (dynamic).

Although dynamic pressure and temperature profiles are not directly used in 3D numerical simulation, the value of this information is when it is analyzed using wellbore simulators. This wellbore simulation helps to further understand the flow mechanism within the wellbore and within the producing formations, which in turn provides the knowledge of the number of production zones and their individual productivities in terms of mass flow and enthalpy at the different wellhead pressures at which the well was flowing when the P/T tools were inside the wellbore. This information is also crucial when doing the evaluation of the pressure at which the different wells shall be operated to draw the maximum of energy at sustainable levels.

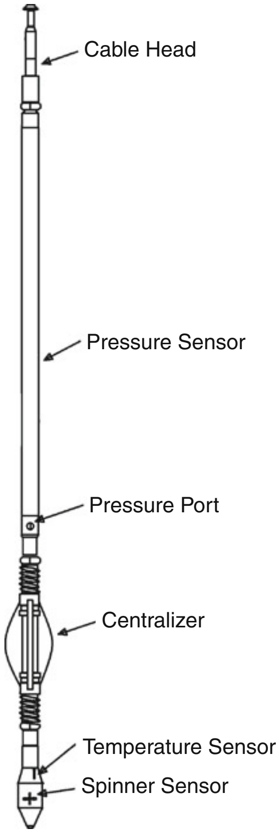


Reservoir Engineering in Geothermal Fields. Figure 2
Example of pressure and temperature logging results

Permeable Zone

To identify the permeable zone(s) in a well, it is necessary not only to characterize the well itself, but also, it is important in the construction of the distribution of the permeable zones within the reservoir. Generally, permeable zones appear when wells encounter fault structures and/or fracture zones associated to fault structures or contact between formation and intrusive rocks or appear when the well taps permeable formations (even not related to fault structures) or when striking a zone of contact between different formations. After the completion of drilling works, PTS (Pressure, Temperature, and Spinner) surveys are carried out to detect the location (depth) of the permeable zones more precisely. PTS surveys are carried out by injecting water into the well; the spinning element of the tool rotates proportionally to the amount of injected water. When the injected water reaches permeable zones, part of the mass flow is lost to the pay zone; therefore, the rotation of the spinning tool changes. The depths at which spinning values change are recorded, thus providing number and depth location of permeable zones. The relative spinning between wellbore locations prior and after the permeable zone is an indicator of the proportional amount of mass lost to the permeable zone (Fig. 3) (please also refer to Permeability).

The spatial positioning of permeable zones in wells contributes to the construction of the geothermal conceptual model, which in turn will be the base of the 3D numerical model. In the 3D numerical model, relatively higher permeability values are given to the elements where permeable zones were identified by either surface exploration (faults, tectonic structure, etc.) or wellbore surveys to represent the distribution of permeability within the geothermal reservoir. Since the simulated fluid flow tends to happen through the permeable elements of the 3D numerical model, the distribution of the permeable zone and the value of permeability are crucial for the modeling work so as to reproduce pressures, temperatures, and flow patterns, as defined in the conceptual model.



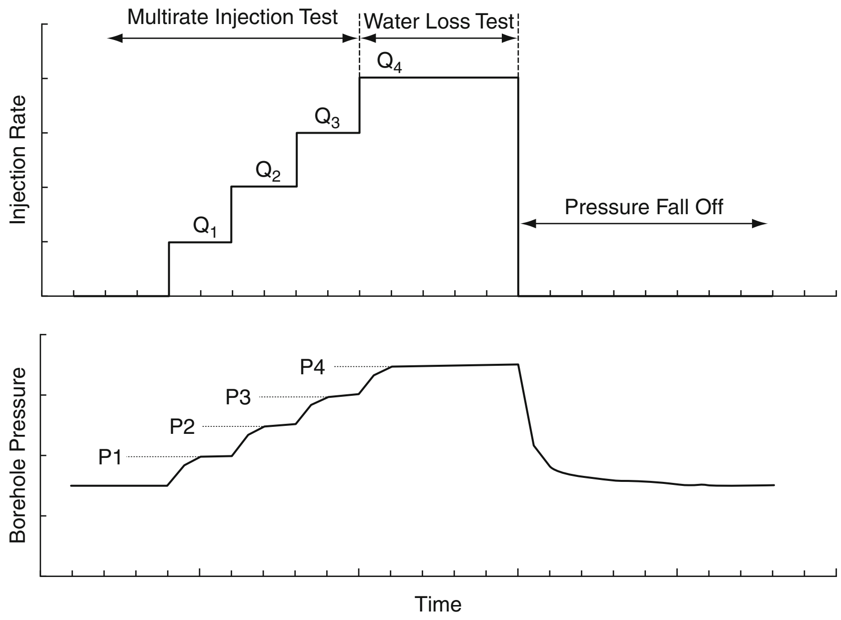
Reservoir Engineering in Geothermal Fields. Figure 3 PTS logging tool

Permeability

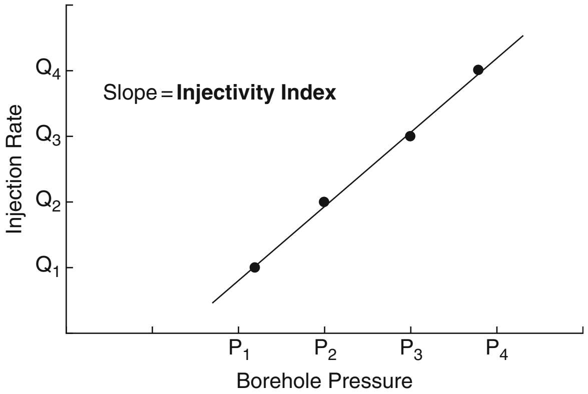
The permeability is a measure of the ability of the underground formations to permit the flow of fluids. A well tapping a permeable zone will also benefit of the capacity of the formation to produce geothermal fluids and the ability of the reservoir to sustain production; therefore, to evaluate the well's permeability is an important factor. In terms of the parameter to be obtained from direct measurements when carrying injection tests, the parameter to be estimated is the permeability–thickness product (the so-called kh value) and not permeability itself. Injection tests generally consist of three tests: multirate injection test, water loss test, and pressure falloff test. The necessary tool for this test is a PT or a PTS downhole tool. The procedure for the multirate injection test is to lower the tool at the depth of the main permeable zone as it was identified during the drilling records or through wellbore PT surveys. Then, water is injected into the well at three or four different injection rates. The injections are sequentially done, and each of the flow rates is maintained during a time in which pressure reading in the PT tool stabilizes. (When only mechanical tools are available, it is advisable to inject each of the rates at the longest time permitted by the water storage, and it is recommendable to use the longest time clock to permit running all injections and other tests in one operation. Refer to Fig. 4.) After the last injection rate, the tool will be run up and down to obtain PT(S) profiles vs. depth during injection (water loss test) which will help to define other permeable zones. After completing the water lost test, the PT(S) tool is left at the depth of the known most important permeable zone, and water pumping is stopped to start the recording the pressure falloff. If mechanical tool is used, the instrument is left for several hours (depending on the clock in use) to assure stabilization of pressure readings upon which the tool will be retrieved from the well. This operation is called pressure falloff test. Using the results of the multirate injection test, the "injectivity index" will be calculated, as shown in Fig. 5. This index expresses the rate of mass the well is able to accept per unit pressure buildup. This value is the simplest crude gauge of the total permeability of the well. The results of the falloff test will be used to estimate the flow capacity such as permeability–thickness products (kh value) of permeable formations surrounding the well. Injectivity index and kh value are very important reservoir parameters to evaluate the productivity and/or injectivity of the well. Figure 6 shows the example of results of the analysis of pressure falloff test. An example of the results of water loss test is shown in Fig. 7.

Discharge Characteristics

The records of the discharged mass flow of all production wells and reinjection mass flow of all reinjection wells are necessary as input data for dynamic numerical simulation. To measure the discharge mass flow rate, adequate equipment should be prepared. The type and size of equipment needed for a "production test" depend on the type of well and reservoir being tested and on the expected maximum flow rate and discharge



Reservoir Engineering in Geothermal Fields. Figure 4 Injection test procedure



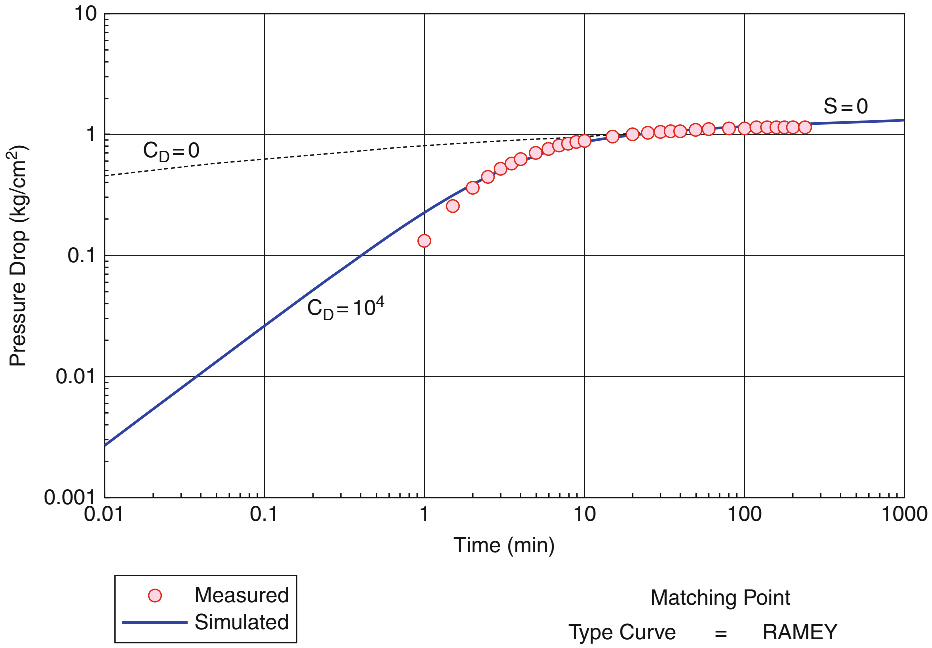
Reservoir Engineering in Geothermal Fields. Figure 5 Injectivity index

enthalpy. In general, one of two kinds of test methods is selected for production test. One is the "separator method," and another is the "lip pressure method."

In the separator method, the discharged fluid, which is, in most case, the mixture of vapor phase and liquid phase, is separated into steam (vapor phase) and brine (liquid phase) at the separator. The separated steam mass flow rate is measured using an orifice plate which is installed in the steam pipeline. The flow rate can be calculated from the differential pressure measured at upstream and downstream of the orifice plate. The brine mass flow rate can be calculated by measuring the water height at the weir and applying the general equation of triangular or rectangular weir.

Figure 8 shows the equipment setting used for the lip pressure method. The empirical equation used by this method relates the enthalpy of the two-phase fluid and the mass flow rate to the lip pressure measured at the end of the discharge pipe, where mass flow and enthalpy are usually unknown parameters. However, using an atmospheric separator and a weir, the mass flow of the water separated at atmospheric conditions can be calculated by measuring the water level in the weir. Considering an isenthalpic expansion and a numerical root-finding procedure, the production enthalpy and the total mass flow (combined steam and water, exclusive of noncondensable gas, at wellhead conditions) can be calculated. The input values are weir height, size of discharge pipe, and thermodynamic properties of water.

To obtain the discharge characteristics of a well, production tests are carried out. The discharge flow rate is continuously measured at least at three or four different wellhead pressures so as to allow enough data



Tool Set Depth	=	1,220 m
Kinetic Viscosity	:	1.56E-07 (m^2/sec)
Permeability-Thickness	=	3.8 (darcy-m)

Matching Point

Type Curve	=	RAMEY
ΔP	=	1
P_D	=	0.65
Δt	=	1
t_D	=	1.8E+04
C_D	=	1.0E+04
S	=	0

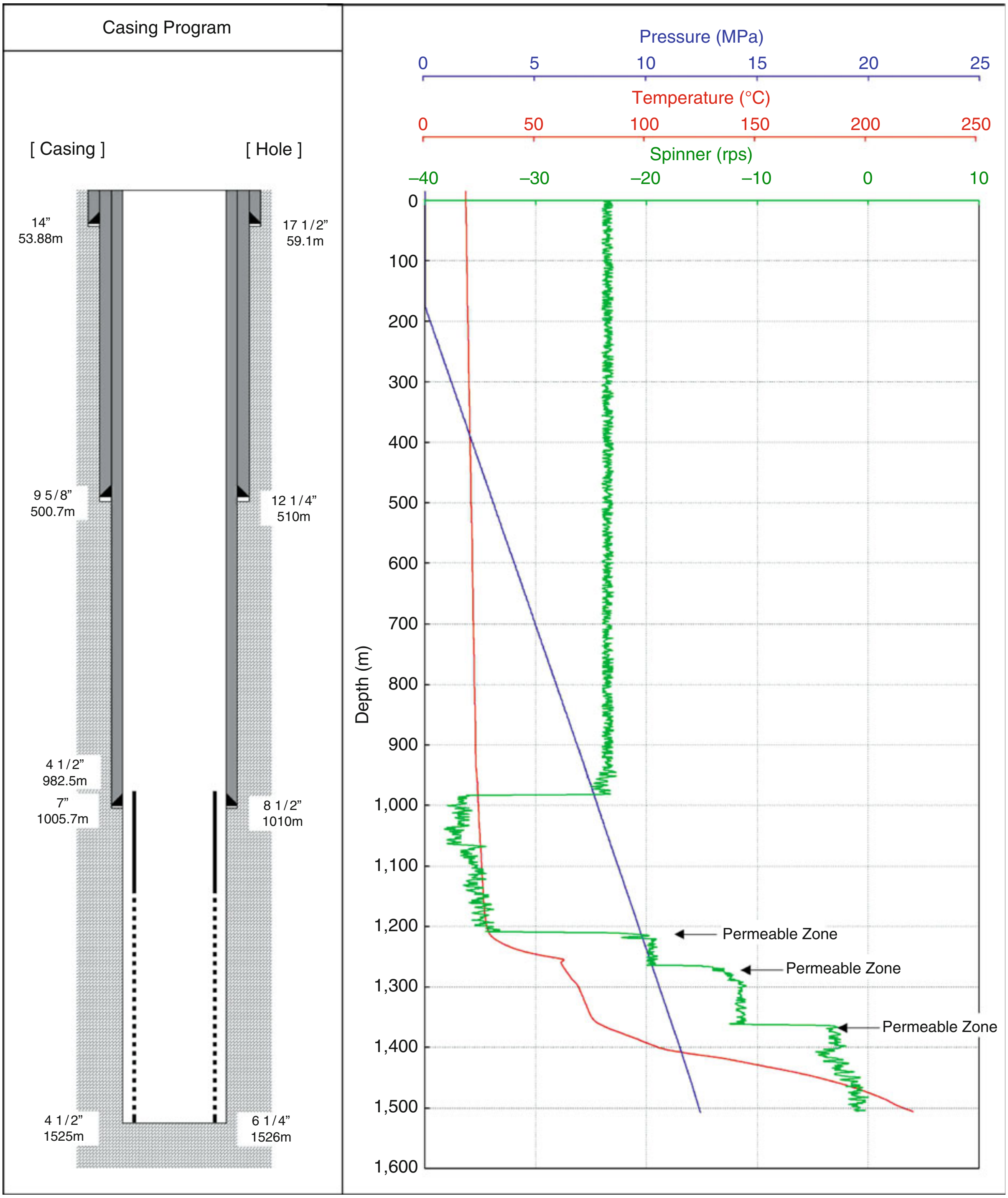
Reservoir Engineering in Geothermal Fields. Figure 6
Example of analysis result of falloff test (curve matching method)

to plot the discharge characteristic curve. To modulate the amount of mass discharge and the wellhead pressure that will permit to plot the discharge characteristics curve (enthalpy, steam, hot water mass flow vs. wellhead pressure), the side valve is opened or throttled. An example of discharge characteristic curve is shown in Fig. 9.

Wellbore Simulation

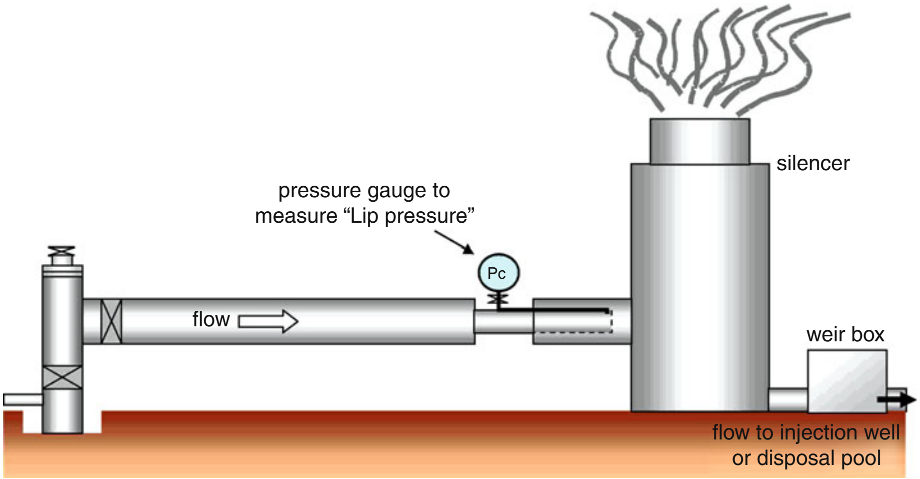
The objective of wellbore simulation is to analyze production data to determine reservoir parameters by repetitive calculation at each of the simulation result to the measured discharge characteristic curve and pressure/temperature profiles. The productivity of the well is mainly controlled by the reservoir pressure and temperature around the well and its permeability. The wellbore simulator has been developed to calculate the relationship between wellhead pressure and steam mass flow rate, water mass flow rate with enthalpy. The main input parameters are feed point depth, well configuration (casing program), reservoir pressure, reservoir temperature (or fluid enthalpy), and permeability–thickness products (kh value). Figure 10 shows wellbore model with the surrounding formation. The reservoir parameters can be estimated by matching the simulation result to the measured discharge characteristic curve. The example of the matching result to the discharge characteristic curve is shown in Fig. 11. The wellbore simulator can also calculate the pressure and temperature profiles vs. depth, and the calculated profiles will be compared to the measured dynamic surveys so as to confirm the enthalpy of the fluid at the point where the permeable area was identified. An example is shown in Fig. 12.

When the 3D numerical simulation is done using a reservoir simulator coupled to a wellbore simulator, the temperature and pressure of each of the element of

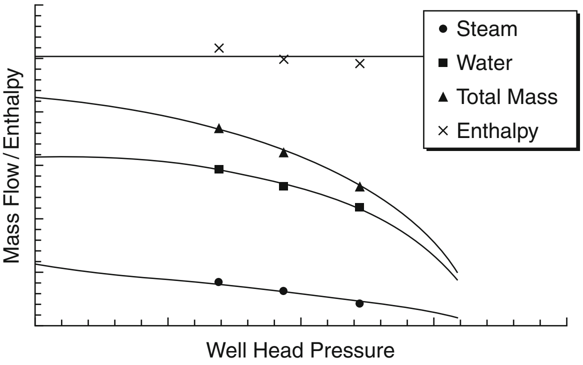


Reservoir Engineering in Geothermal Fields. Figure 7

Example of water loss test using PTS tool



Reservoir Engineering in Geothermal Fields. Figure 8
Schematic equipment setting for the lip pressure method



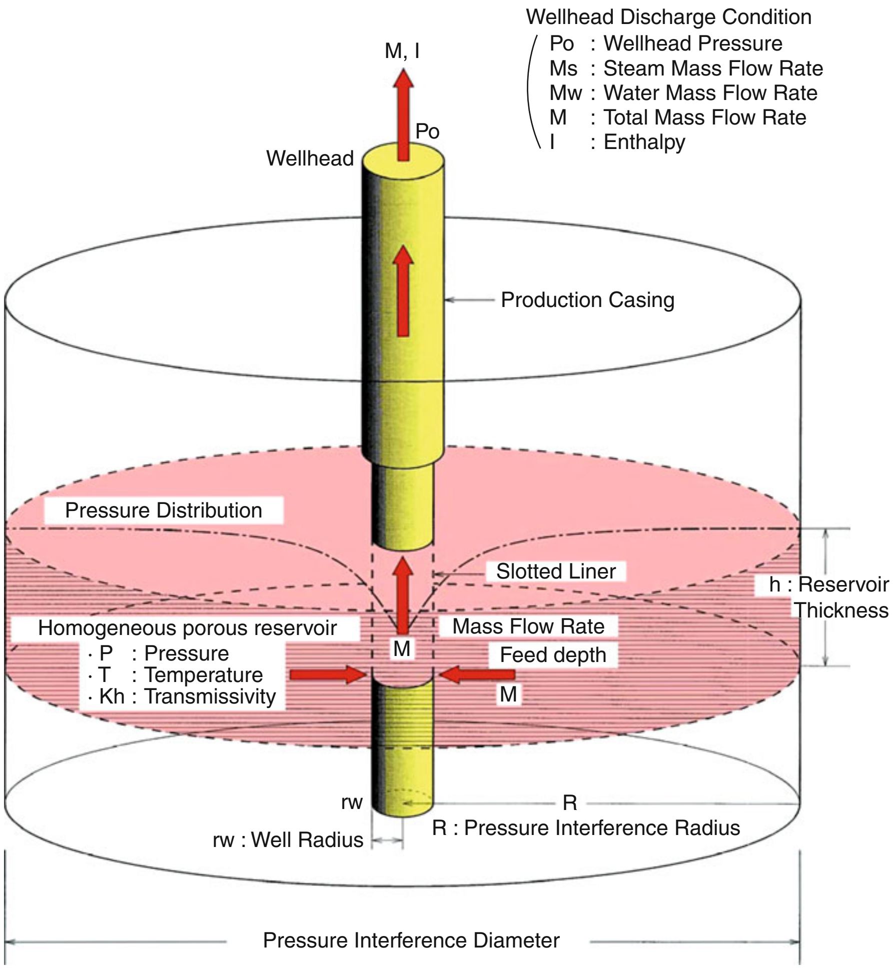
Reservoir Engineering in Geothermal Fields. Figure 9
Example of discharge characteristic curve

the numerical model that corresponds to a permeable zone identified in a well should match the temperature and pressure calculated by the wellbore simulation analysis after match is attained between calculated and measured production characteristics curves. Once the numerical model of the reservoir is calibrated (definition of the permeability characteristics of the well), the future change of discharge characteristics can be calculated from the values of pressure, temperature, and/or enthalpy predicted by the 3D numerical simulation for the elements containing the permeable zone of the wells. Nevertheless, in this calculation, the kh value is assumed constant during all the time of simulation period. Actually, kh values increase due to the continuous discharge or can decrease due to scale deposition. This is an issue necessary to be challenged for a more realistic forecasting simulation.

Interference Among Wells

During the discharge test, the reservoir pressure (and sometimes temperature also) is monitored at the observation well. The objective is to verify whether the monitoring pressure changes are not due to the production and/or reinjection of the geothermal fluid. For the monitoring, in general, capillary tubing system is used as pressure monitoring tool. The capillary tubing system consists of pressure chamber, capillary tube, helium gas vessel, and transducer with data recording system. The pressure chamber is set at the depth where it is detected as permeable zone and then filled with helium gas. The change in helium gas pressure in the pressure chamber with capillary tube, which indicates reservoir pressure, is transferred to a data recording system through the transducer. It is necessary to start pressure monitoring before the discharge test because there might be some natural trend of the pressure change which is not related to the testing activity. Sometimes tidal fluctuation in pressure can be observed.

The observed pressure data is examined considering the actual discharge and reinjection record. If there is a relationship between the mass extraction/injection and the change in monitoring pressure, it indicates the



Reservoir Engineering in Geothermal Fields. Figure 10
Wellbore simulation model with the surrounding formation

hydraulic connectivity and mutual interference between the wells. The analysis of this data provides reservoir properties, such as the permeability–thickness products (kh) and reservoir boundary. This information is very useful to construct the 3D numerical model, and also the observed pressure change should be one of the guidance to calibrate the numerical model.

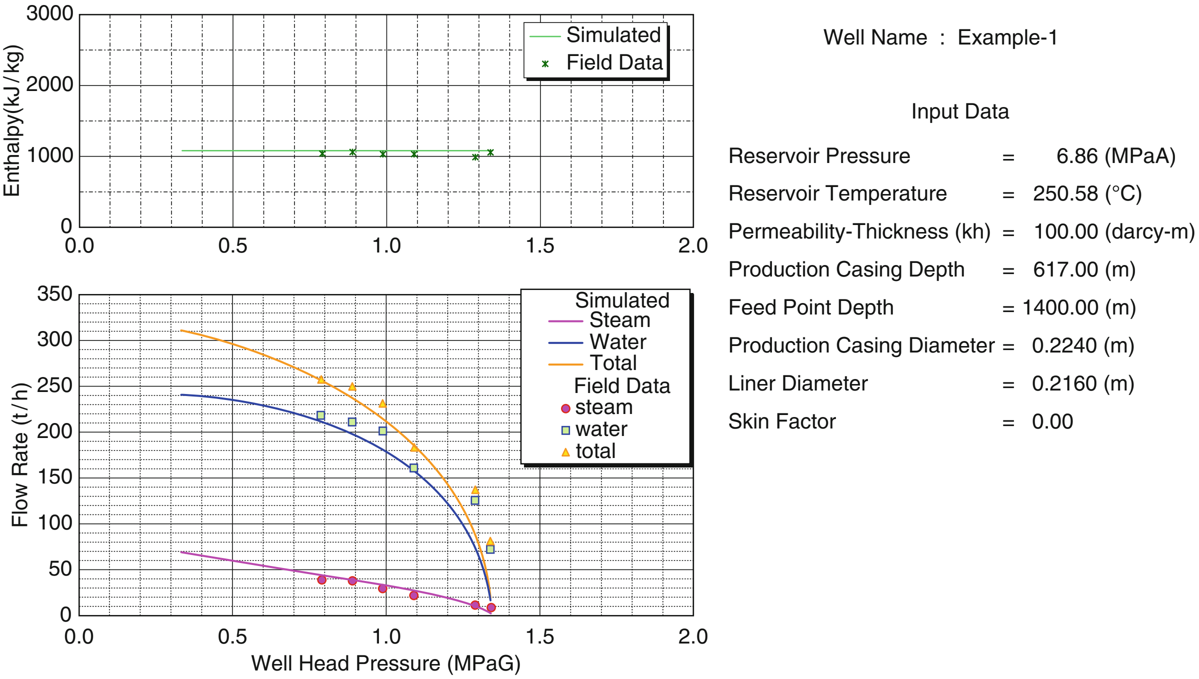
Tracer Test

Tracer testing, which is used to study flow paths and quantify fluid flow in hydrological systems, is probably the most efficient tool to evaluate the direct communication between reinjection well and production well. Tracer tests involve injecting a chemical tracer into a hydrological system and monitoring its recovery through time at various observation points. The results are, consequently, used to study flow paths and quantify fluid flow.

When designing a tracer test, the following aspects must be considered carefully:

1. What tracer to select
2. The amount of tracer to inject
3. The sampling plan to follow (sampling points and frequency)

The tracer selected needs to meet a few criteria: (1) It should not be present in the reservoir (or at a constant concentration much lower than the expected tracer concentration). (2) It should not react with or



Well Name : Example-1

Input Data

Reservoir Pressure	=	6.86 (MPaA)
Reservoir Temperature	=	250.58 (°C)
Permeability-Thickness (kh)	=	100.00 (darcy-m)
Production Casing Depth	=	617.00 (m)
Feed Point Depth	=	1400.00 (m)
Production Casing Diameter	=	0.2240 (m)
Liner Diameter	=	0.2160 (m)
Skin Factor	=	0.00

Reservoir Engineering in Geothermal Fields. Figure 11
Example of the matching result to the production characteristic curve

absorb to the reservoir rocks. (3) It should be easy (fast/inexpensive) to analyze. The following are the tracers most commonly used in geothermal applications:

1. Radioactive tracers like iodide-125 (^{125}I), iodide-131 (^{131}I), tritium (^{3}H), etc.
2. Fluorescent dyes such as sodium fluorescein
3. Chemical tracers such as iodide, bromide, etc.

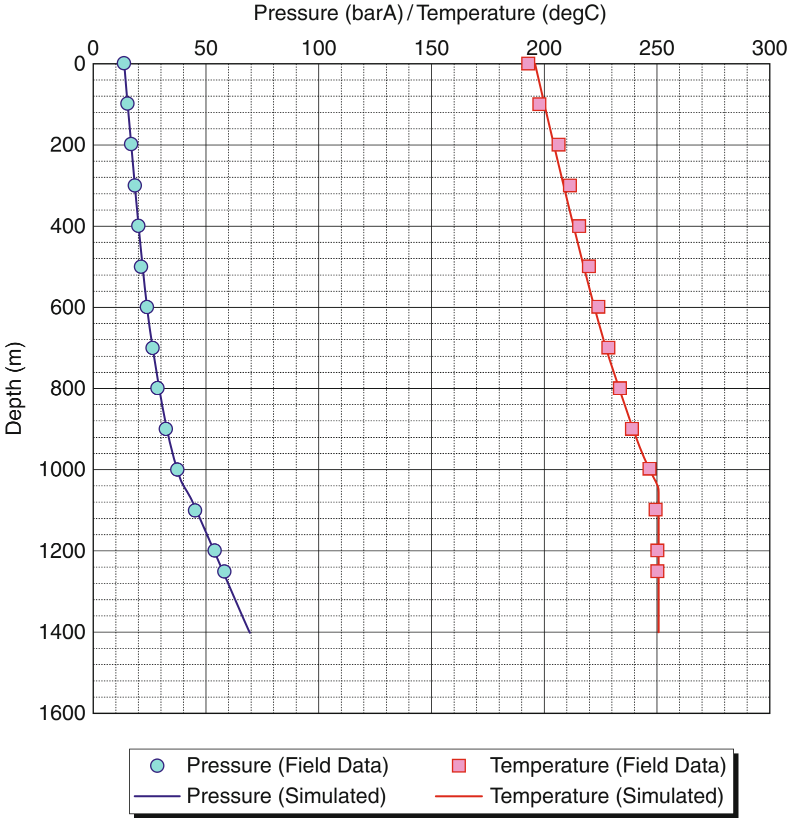
Sodium fluorescein has been used successfully in numerous geothermal fields. Fluorescein has the advantage of being absent in natural hydrological systems. It may also be detected at very low levels of concentration. Furthermore, the concentration of fluorescein is measured very easily, it being a fluorescent dye. The main disadvantage in using fluorescein is that it decays at high temperatures.

After a suitable tracer has been selected, the mass of tracer to inject needs to be determined. This is always difficult to determine beforehand but depends on several factors:

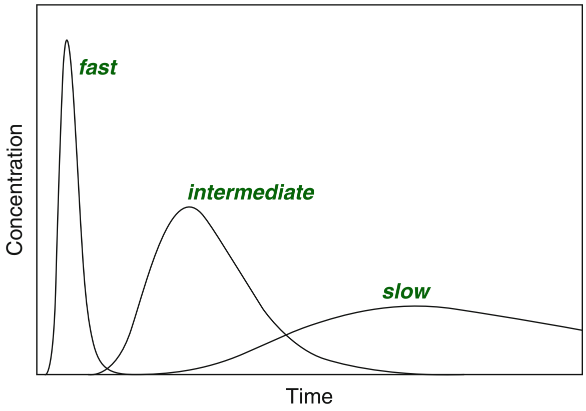
1. Detection limit
2. Tracer background (if any)
3. Injection rate
4. Production rate and how many wells are involved
5. Distances involved
6. Return rate anticipated (slow/fast)

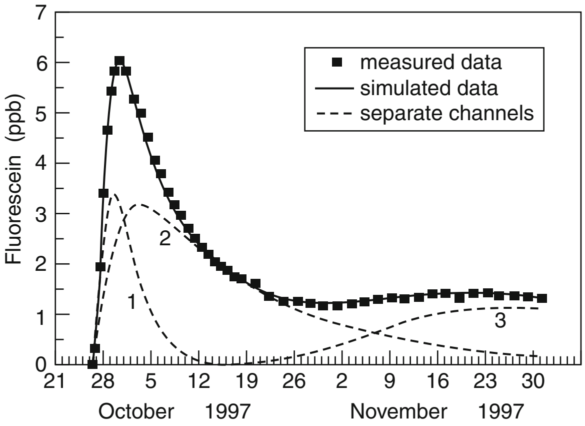
The length of a tracer test depends on local reservoir conditions and distances between wells involved, which control the fluid flow pattern in the reservoir (Fig. 13). They usually last from a few weeks to months or even years. When distances are long and/or fluid flow is slow, tracer tests must be expected to be quite long. The length is preferably not determined beforehand, however, since the rate of return is hard to forecast. Once a sufficiently good data set has been obtained, a tracer test may be terminated.

An example of analysis of tracer test data is shown in Fig. 14. Three separate flow channels were used in the analysis, which are assumed to connect the different feed zones of the injection and production wells. The properties of the channels are different from each other. This kind of information can contribute to the construction of 3D numerical model, although such



Reservoir Engineering in Geothermal Fields. Figure 12
Example of the matching result to the pressure and temperature profiles





Reservoir Engineering in Geothermal Fields. Figure 13
Typical fast, intermediate, and slow tracer return profiles [1]

Reservoir Engineering in Geothermal Fields. Figure 14
An example of analysis of tracer test data [1]

detailed path setting and reproduction is not easy. However, the trial to reproduce the tracer return in 3D numerical simulation has already been carried out.

Historical Data

After the development of the geothermal power plant, the reservoir condition changes with time because of the extraction and reinjection of geothermal fluid. Thus, the historical data of following parameters is required for the 3D reservoir simulation.

Mass Production and Reinjection The historical data of mass production and reinjection is absolutely imperative for the reservoir management. It is the most basic data to evaluate the reservoir condition and behavior and also to maintain the output of the geothermal power plant. For example, if the mass flow from a production well declines, study for the possible reason (pressure decline, scaling problem, impact of cold reinjection, etc.) is required, and proper countermeasures should be applied. In any case, the historical data of mass production and reinjection is absolutely important. In the 3D numerical simulation, this data is used as input to the simulator to calculate the change of reservoir conditions.

Enthalpy of Each Well Similarly, fluid enthalpy produced from each well should be monitored. If the fluid enthalpy tends to decline, it indicates the cooling of the reservoir, and thus, the negative impact by the cold reinjection should be suspected. On the other hand, if the fluid enthalpy tends to increase, the reservoir might be drying, and the reservoir pressure drawdown might be serious. In this case, overextraction of the geothermal fluid is suspected. Thus, it is necessary to monitor the fluid enthalpy of each well for the stable operation of the reservoir. The historical enthalpy data can be used as matching target for the dynamic calibration of the 3D numerical model.

Pressure and Temperature If the well out of use is available, it can be used as monitoring well. Pressure and temperature monitoring system (please refer to Interference Among Wells) can be installed to the well. This information is very useful to observe the reservoir behavior and can be a good matching target for the dynamic calibration of the 3D numerical model.

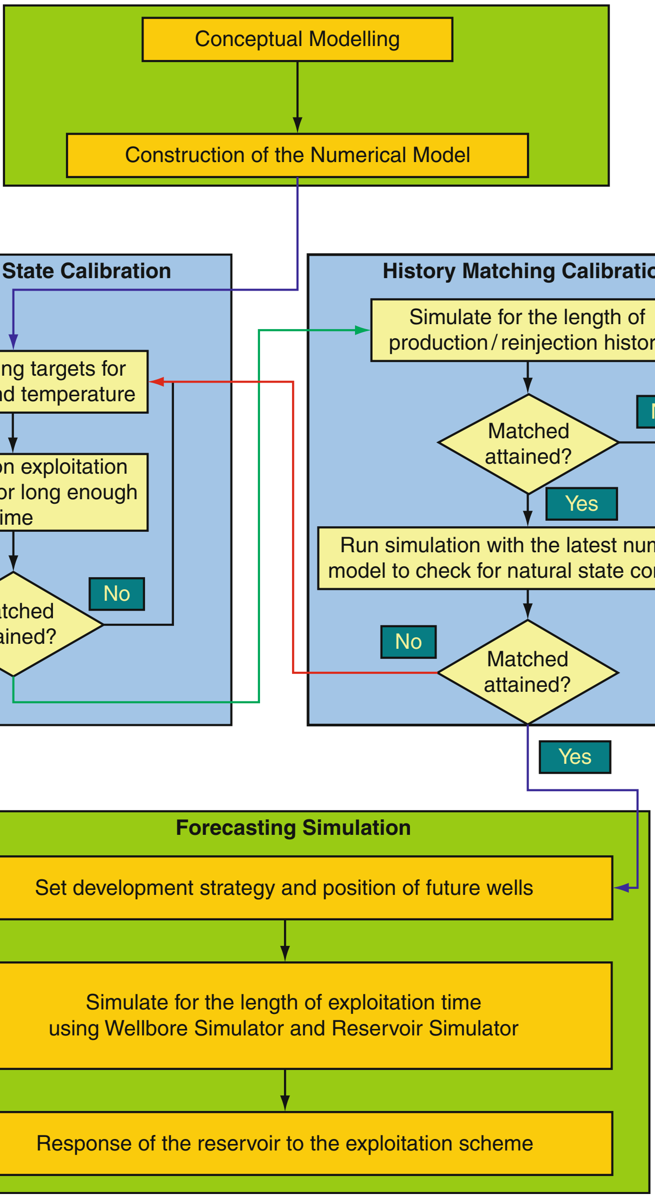
Numerical Simulation

Work Flow

Numerical simulation is a computational work on the reservoir numerical model, which is the most popular technology to quantitatively evaluate geothermal resource potential and also optimize the strategic plan of the steam field development and/or its management. In general, the work of numerical simulation consists of three steps: model establishment, model calibrations, and forecasting (Fig. 15).

Establishment of Reservoir Numerical Model

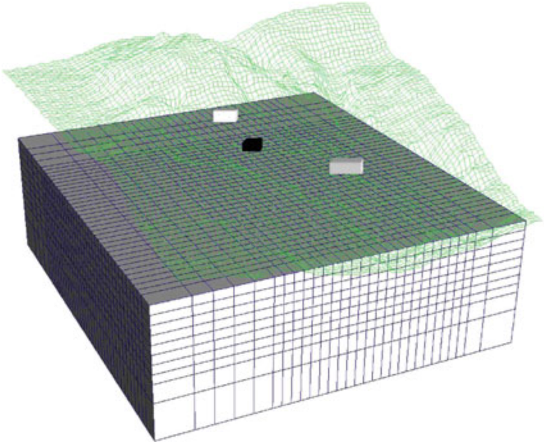
Once the reservoir conceptual model is made, the next step is to establish the reservoir numerical model. The most important thing for establishing the numerical model is that it should coincide with the conceptual model in terms of the structures, heat source(s), and fluid flows in the reservoir. For example, even if the numerical model reasonably explains distributions of the pressure and temperature in the reservoir, ignoring the reservoir conceptual model, it should be revised until the numerical model is harmonized with the conceptual model because the conceptual model is the base of the numerical model. After careful considerations of the conceptual model, a grid is generated to discretize the geothermal system in three dimensions (Fig. 16). The simulation area should cover all the geothermal reservoirs and also their surrounding areas where are, as it were, buffer zones to mitigate impacts of boundary conditions on insides of the numerical model. The boundary conditions define the calculation conditions at all the boundaries at top, lateral sides, and bottom of the numerical model. The boundary conditions are unique concept of the numerical model and necessary to solve equations of mass and energy balances in the numerical model. In areas where reservoir information is available or where the well density is high, the grid blocks should be small. Conversely, in areas where no reservoir information is readily available or where the well density is low, the grid blocks are larger (Fig. 17). In general, the area of the numerical model will range from several square kilometers to several 10 km^2, depending on the reservoir conceptual model. In the vertical dimension, the model should extend to the basement rocks or to the


Conceptual Modelling
Construction of the Numerical Model
Natural State Calibration
Set matching targets for pressure and temperature
Perform non exploitation simulation for long enough time
Matched attained?
No
Yes
History Matching Calibration
Simulate for the length of production / reinjection history
Matched attained?
No
Yes
Run simulation with the latest numerical model to check for natural state conditions
No
Matched attained?
Yes
Forecasting Simulation
Set development strategy and position of future wells
Simulate for the length of exploitation time using Wellbore Simulator and Reservoir Simulator
Response of the reservoir to the exploitation scheme

Reservoir Engineering in Geothermal Fields. Figure 15
Flowchart of reservoir simulation works

depth that is deep enough to mitigate the impacts of boundary conditions at the bottom of the model on the reservoir depths. Therefore, in many cases, the overall thickness of the numerical model will be a few kilometers and is subdivided into the number of layers, depending on the sectional information such as distributions of the geological formations, cap rocks, reservoirs, and basement rocks. For example, the thickness of each layer will be a few 100 m. Some grid blocks corresponding to the elevations close to the top of the model are eliminated, depending on the topographic elevation.

Once the grid system is constructed, the required grid block parameters are entered. Rock properties required for the reservoir simulation are density, porosity, permeability, specific heat, and thermal



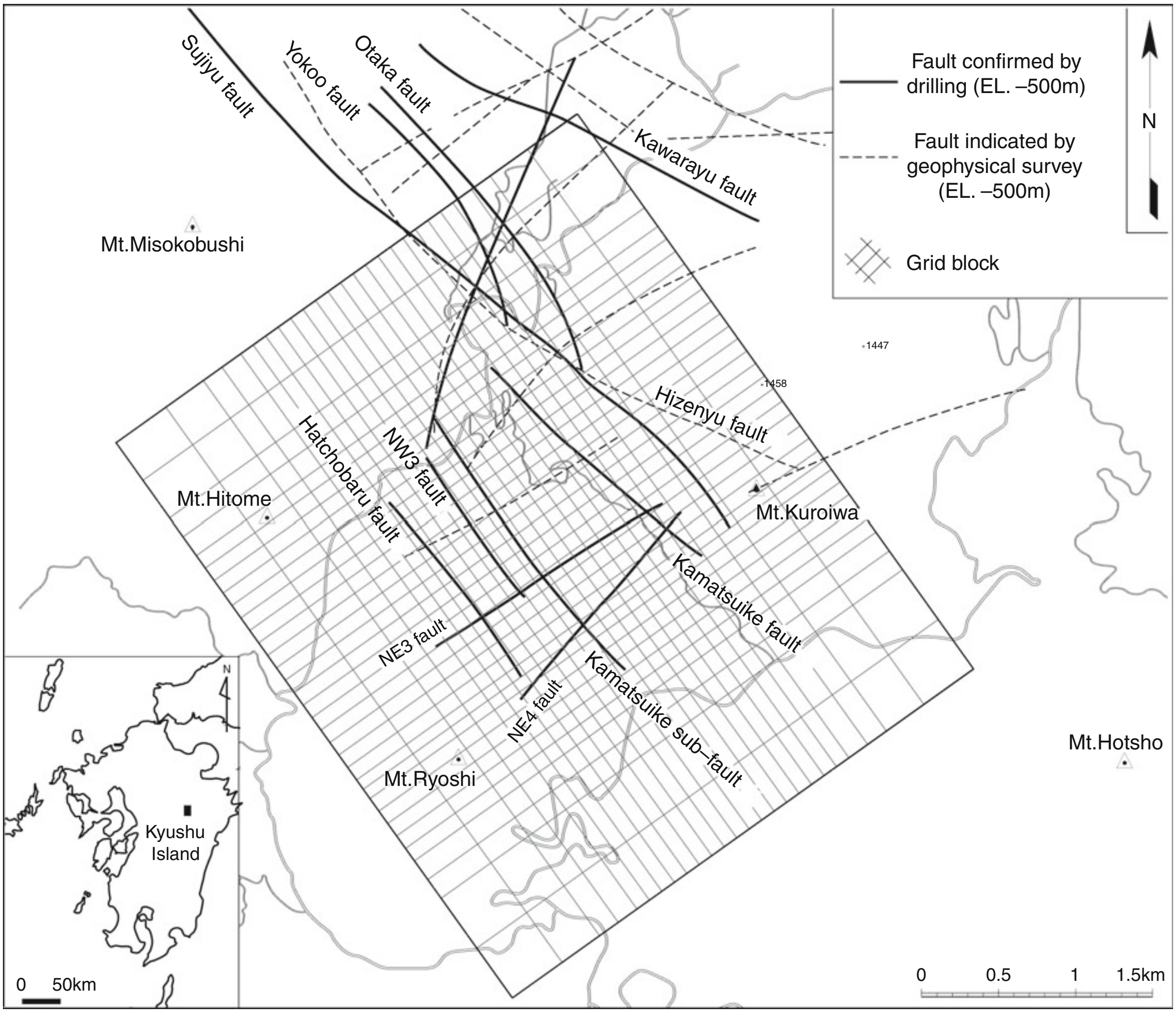
Reservoir Engineering in Geothermal Fields. Figure 16 Three-dimensional view of a reservoir simulation grid model

conductivity of each rock. The rock properties should be set for all grid blocks (Fig. 18). The most important parameter is permeability because permeability of the rocks usually has a wide range, depending on the rock type and structure and depending also on the control flow direction and mass flow rate of the fluids in the reservoirs. In order to estimate the permeability distribution, resistivity distributions are good reference because low-resistivity regions are considered to be low-permeability regions that play the role of cap rocks. Cap rocks limit the reservoir extent by preventing high-temperature fluids from flowing out into the surroundings. Thus, reservoirs with high-temperature fluids will be formed beneath the cap rocks, meaning that the distribution of low resistivity that is correlated with cap rocks may represent the reservoir distribution. The basic distributions of rock properties were examined, with a consideration of the resistivity distribution, geological formations, and fault structures in the geothermal field. Then both rock properties and their distribution were repeatedly modified by trial and error until the numerical model could reasonably explain the natural-state conditions in the reservoirs.

Calibration of the Reservoir Numerical Model

In order to make the reservoir numerical model reliable and available for predictions of reservoir performance under exploitation, a two-step calibration of the reservoir numerical model should be made. The first-step calibration is a natural-state simulation in which the reservoir numerical model should represent the steady state conditions of the reservoir under preproduction. The next step calibration is a history matching in which the reservoir numerical model satisfactorily representing the natural-state conditions should also represent the dynamic conditions of the reservoir during the exploitation. In these calibrations, the judgment of appropriateness of the numerical model should be done by comparison between the measured values and the corresponding computed values of each step simulation. When the reservoir numerical satisfies the two-step calibration, it is judged that the model can be applied to the forecasting simulation.

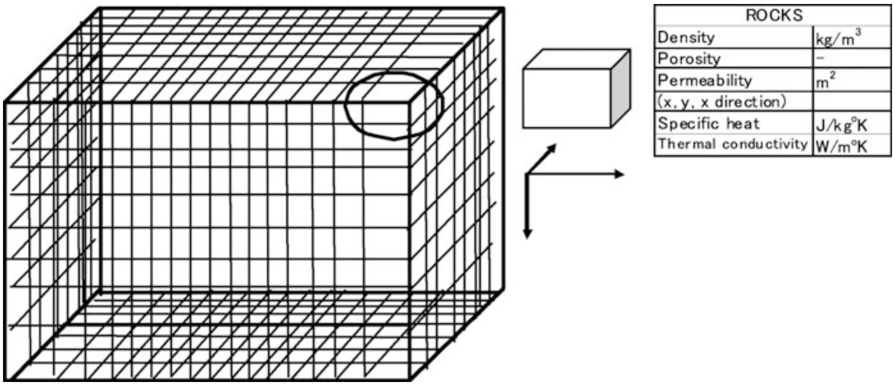
Natural-State Simulation Geothermal systems evolve over geologic time, during which the thermodynamic and hydrodynamic conditions in the system move toward a dynamic equilibrium. The rate of this change is exceedingly small compared to the changes induced during the exploitation of such a system. Therefore, for all practical purposes, undeveloped geothermal systems are considered to be in a quasi-steady state. The main objective of the natural-state calibration is to verify the distributions of temperature, pressure, and steam faction in the reservoir and the heat/mass flow aspects of the model. Therefore, more detailed distributions are needed to be specified for the permeability and thermal conductivity of the rock in order to match the subsurface temperature and pressure distribution. Once the natural-state calculation starts, high-temperature fluids will begin to flow upward due to buoyancy along the higher permeable structures or formations from the zones where the boundary conditions are set as upflow zones or heat sources at the bottom of the model. The fluids will decrease temperature and buoyancy due to contamination with the surrounding fluids with lower temperature and then begin to flow downward after reaching the depth of cap rocks where the fluid flow is constraint due to small permeability. The fluids will flow down to the basement rocks and recover temperature and buoyancy again at the depth, and then the natural convection will eventually occur (Fig. 19). A set of initial values for the petrophysical properties of the materials is selected,



Reservoir Engineering in Geothermal Fields. Figure 17
Plan view of the grid used to model the Hatchobaru geothermal field [4]

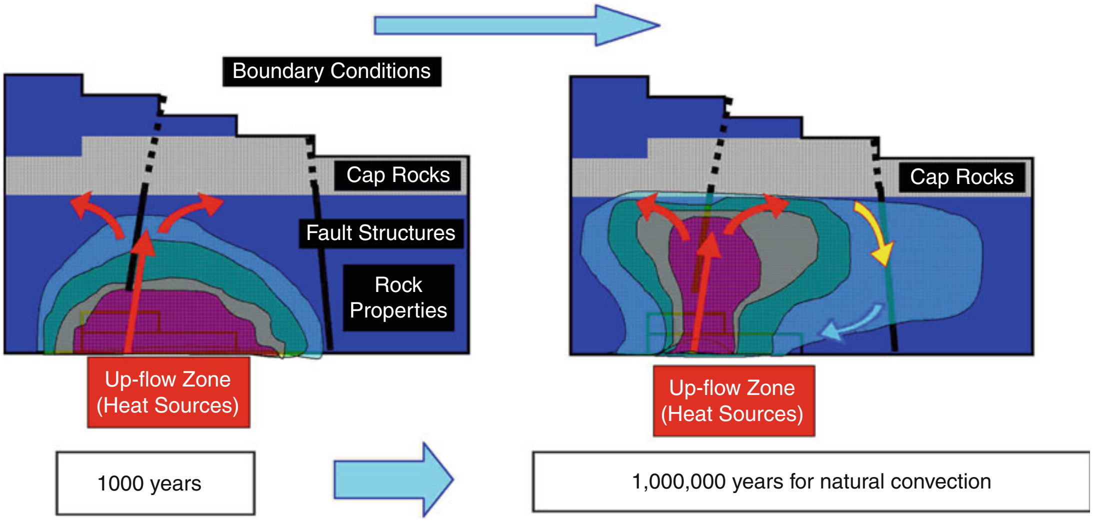
and the model is allowed to run to a quasi-steady state condition, which takes, in simulated time terms, an order of more than 10^5 years. When the model fails to reach a quasi-steady state, it is because the numerical model is unstable or not realistic, and the input parameters have to be modified accordingly. After a lot of iteration, a temperature and pressure match between measured and simulated will be obtained (Figs. 20 and 21).

History Matching When historical data of mass flow rates of production and injection wells is available and also when there is availability of pressure and temperature changes over time during the production monitored by observation wells, the history matching can be done as the second calibration step of the numerical model following the natural-state simulation. It is a basic rule of the reservoir simulation that the numerical model always employs the same distributions of rock properties and boundary conditions which were imposed in the preceding the natural-state simulation. In the history matching, the simulator is provided with the mass extraction history as input data and reservoir response history such as pressure and temperature changes over time as reference data. The accuracy of the numerical model depends on the results



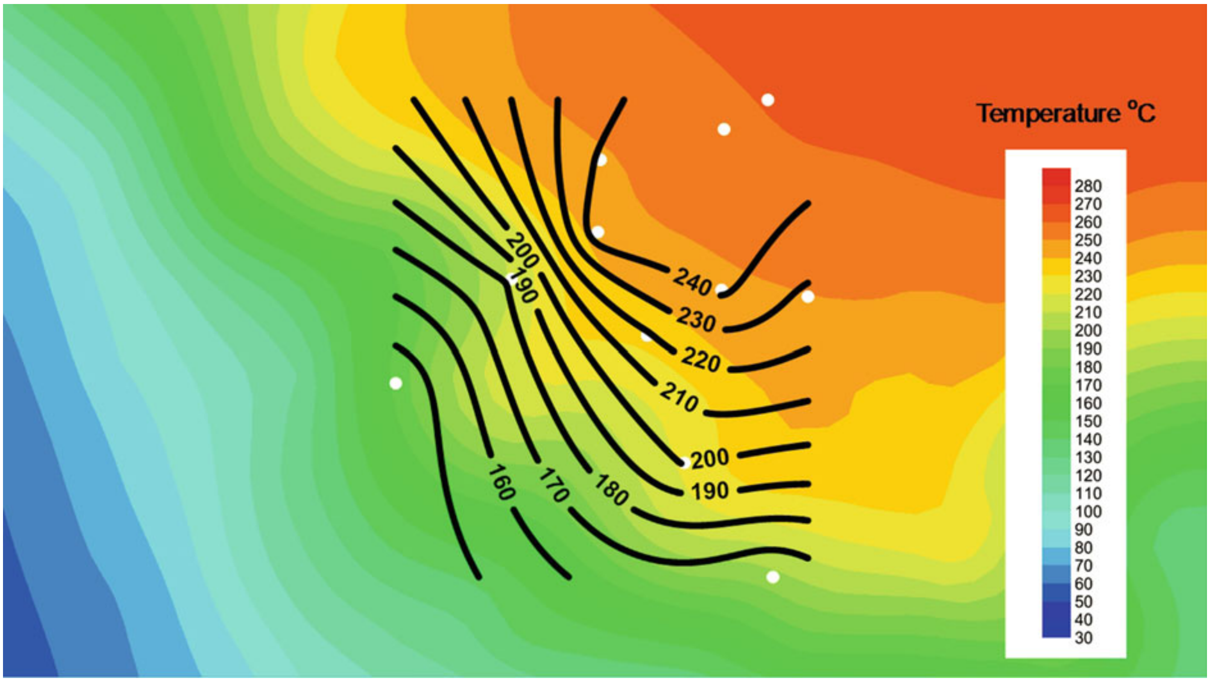
ROCKS	
Density	kg/m^3
Porosity	–
Permeability	m^2
(x, y, x direction)	
Specific heat	J/kg°K
Thermal conductivity	W/m°K

Reservoir Engineering in Geothermal Fields. Figure 18
Rock properties assigned to an element of the reservoir numerical model

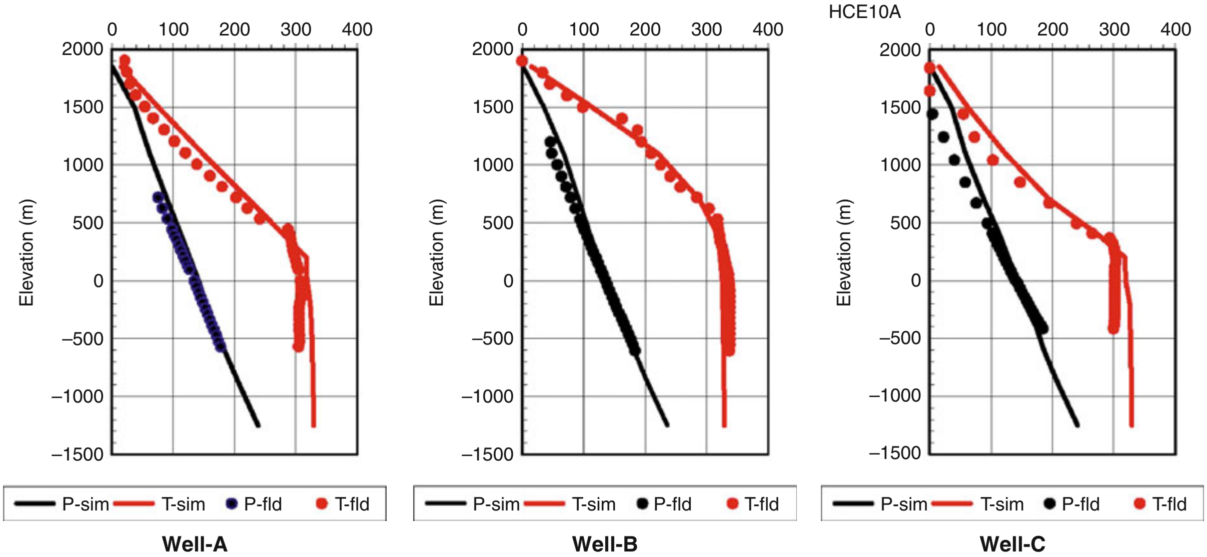

Boundary Conditions
Cap Rocks
Fault Structures
Rock Properties
Up-flow Zone (Heat Sources)
Cap Rocks
Up-flow Zone (Heat Sources)
1000 years
1,000,000 years for natural convection

Reservoir Engineering in Geothermal Fields. Figure 19
Simulated natural convection of geothermal fluids in natural state

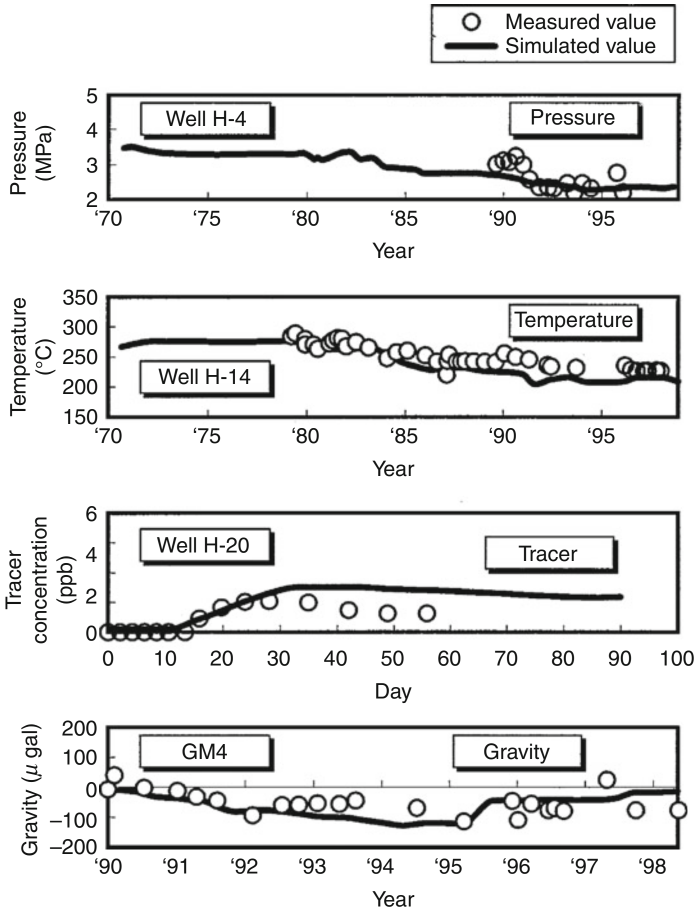
of both natural-state simulation and history matching. Therefore, quality and quantity of the measured data for the reference of the matching is very important. The longer and much more the term and items of reservoir monitoring during exploitation or plant operation are, the higher the accuracy of the numerical model is. For example, it is possible to use as reference the monitoring data of pressure, temperature, and gravity changes over time, and also the result of tracer return during plant operation [2]. In the Hatchobaru geothermal field, Japan, partial success in satisfactorily matching results with monitoring data of pressure, temperature, gravity changes, and tracer behavior was obtained (Fig. 22). In the history matching, satisfied matches to such various kinds of measured data will increase confidence of the numerical model because a lot of measured data to be referenced takes a role of providing considerable constraint to the numerical model. If significant disagreement exists between the computed results and field observations, appropriate adjustments should be made in the parameters in the numerical model, such as permeability distribution



Reservoir Engineering in Geothermal Fields. Figure 20
Matching of temperature distributions in the natural-state simulation (black: derived from measured values; colored: calculated values)



Reservoir Engineering in Geothermal Fields. Figure 21
Matching of wellbore static temperature and pressure profiles in natural-state simulation



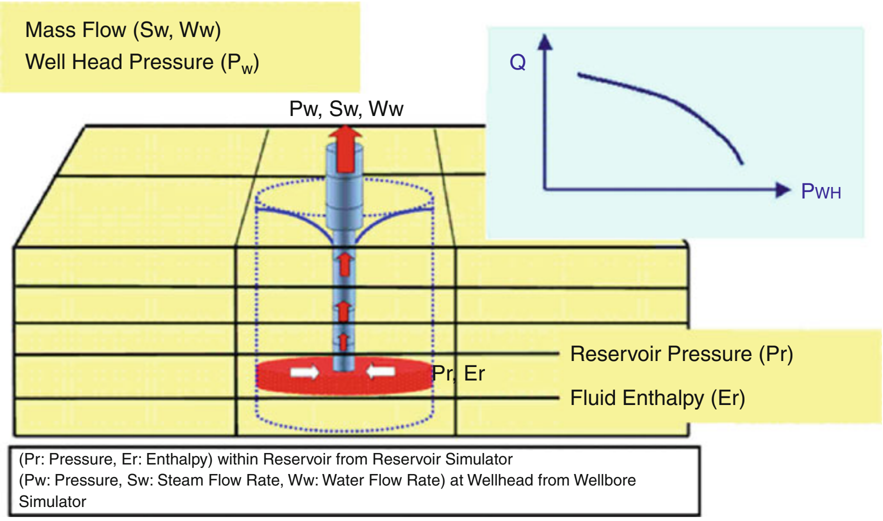
Reservoir Engineering in Geothermal Fields. Figure 22
Matching to reservoir monitoring data during the operation of the Hatchobaru power plant, Japan

and boundary conditions, and the entire process should be repeated, starting with the recalculation of the natural state, until satisfied history matching can be obtained. Accordingly, the entire matching process for the model calibration usually needs to be repeated numerous times to achieve satisfactory agreement between the computed results and the available field measurements.

Forecast Simulation

Once the history matching process is complete, the calibrated model can be used to make predictions of the future behavior of the geothermal field under the hypothetical development scenarios. Results of the forecast simulation will clarify degree of sustainable power development in the geothermal field, which can be defined as a geothermal resource potential.

In the forecast simulation, the development scenarios, including type and capacity of power plant (required steam and water production rate), location of both production and reinjection wells, and future well drilling plan, have to be assumed. Along the hypothetical scenario(s), the reservoir response will be calculated for long-term operation period (30 years, for example). The forecast simulation is recommended to be conducted by directly coupling the reservoir simulator with the wellbore simulator to have more accurate future performance of the production wells [3, 4]. The wellbore simulator calculates the mass flow rate and power output of each production well at the given wellhead pressure on the basis of the reservoir


Mass Flow (Sw, Ww)
Well Head Pressure (P_w)
Pw, Sw, Ww
Q
PWH
Pr, Er
Reservoir Pressure (Pr)
Fluid Enthalpy (Er)
(Pr: Pressure, Er: Enthalpy) within Reservoir from Reservoir Simulator
(Pw: Pressure, Sw: Steam Flow Rate, Ww: Water Flow Rate) at Wellhead from Wellbore Simulator

Reservoir Engineering in Geothermal Fields. Figure 23
Concept of coupling reservoir simulator and wellbore simulator

conditions that are calculated by the reservoir simulator (Fig. 23) [5–9]. Accordingly, the direct coupling method of a reservoir simulator and a wellbore simulator provides the predictions of the changes of power output over time of not only each production well but also the power plant.

In many geothermal fields in the world, it is reported that decline of the power output of geothermal power plant occurred, more or less, after commissioning the plant operation. The reasons of the power decline are case-by-case and field-by-field; however, they are generally classified into three categories below:

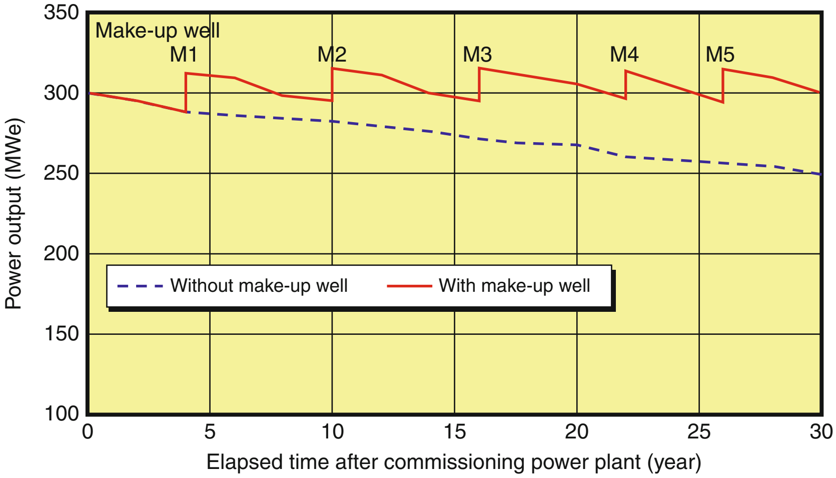
1. Pressure drawdown in the production zones due to interferences between wells or overproduction
2. Cooling in the production zones due to migration of injected brine or local water
3. Decrease of the permeability in the production zones or decrease of useful diameter of pipelines due to scale deposition

The above reasons can be avoided or mitigated by optimizing well allocation in the field or optimizing wellhead pressure. These optimizations can be done through the forecast simulation, by employing the best scenario among the hypothetical scenarios relating to well allocation and wellhead pressure. For example, the pressure drawdown and cooling in the production zones will be avoided or mitigated by keeping enough space between production wells and also between production and reinjection zones [10]. In addition, the scaling problem may be mitigated by setting higher wellhead pressure of production wells, although their productivity may decrease in accordance with higher wellhead pressure. Therefore, optimum wellhead pressure should be discussed in terms of the well productivity and mitigation of the scaling.

In order to maintain the rated output of the power plant through the plant operation life, additional drilling of makeup wells will be required at appropriate timing. The forecast simulation clarifies the number of wells required and appropriate timing of its drilling (Fig. 24). After running the several different scenarios and comparing the results, the most optimum development scenario can be decided.

Sensitivity Study

Sensitivity studies are methods to validate the certainty and uniqueness of the reservoir numerical model and compensate its uncertainty. Sensitivity studies are done by changing the parameters set in the model and then comparing their forecasted results with those of the base model to check how some of the parameters set in the model affect the prediction results. A simple procedure of the sensitivity studies is to change the parameters of the

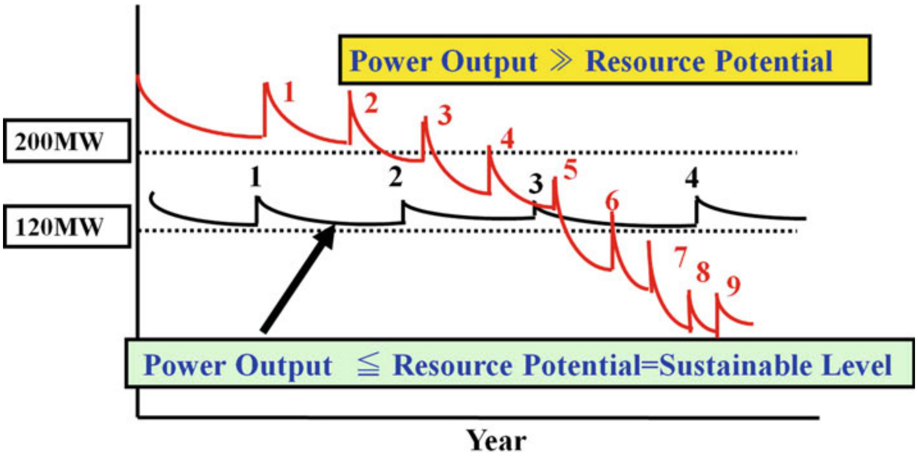


Reservoir Engineering in Geothermal Fields. Figure 24
Forecast simulation to predict the number and timing of commissioning of makeup production wells

numerical model selected and then conduct the forecast simulation to check their influences on the predicted results. For example, if the permeability values of the main productive faults in the numerical model change by 10%, the results forecasted will be changed compared to those of the base model. If they are drastically changed, it suggests that the permeability of the main productive faults is a key factor to determine the model accuracy as well as results of predictions. More complex procedure of the sensitivity study is to try to find other reasonable combination of the parameter settings of the numerical model and then compare the forecasting results to those of the base model. Even if the parameters of the base model are changed, some cases may satisfy the results of both natural-state simulation and history matching. Especially in the early exploitation stage, it will be very difficult to establish a unique numerical model due to poor measured data for the reservoir. In such cases, all the models that indicate fairly good matching in the natural-state simulation and history matching are available for the forecast simulation. Therefore, these models may show some range for the predictions of the decline of power output and/or number of makeup wells required from the conservative to optimistic values. This range can be made small when data of the reservoir is accumulated, which means that the reservoir model becomes more precise and the forecasts becomes more accurate as new data becomes available.

Resource Evaluation

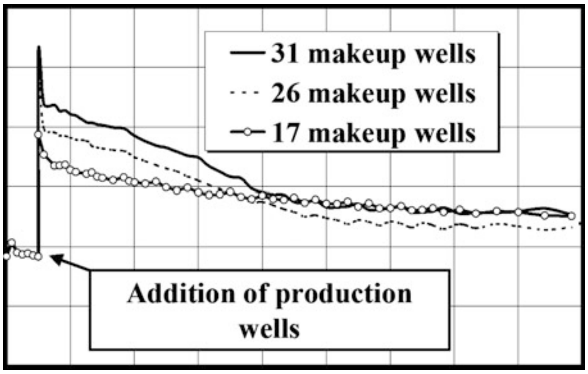
As a general concept, a power plant should maintain its rated power output throughout the plant operation life of around 30 years. The power output of the geothermal power plant should be designed based on its resource potential that sustains continuous provision of the geothermal fluids required to the power plant from the geothermal reservoir(s) through production wells during the plant operation. Taking a balance of fluid production and provision in the reservoir is most important in terms of mass and energy. If the balance is successfully taken, the rated power output will be easily kept throughout the plant operation life. In such case, no or only a small number of makeup wells may be required. On the contrary, if mass or energy is produced significantly beyond the resource potential of the geothermal reservoir, it will be difficult to maintain the rated power output, even if makeup wells are added, and then the power output will be eventually declined (Fig. 25). Therefore, evaluation of resource potential is indispensable for optimizing not only the design of geothermal development but also the reservoir management plan for the stable power plant operation. As an effective method, a case study


Power Output ≫ Resource Potential
200MW
120MW
1
2
3
4
5
6
7
8
9
1
2
3
4
Power Output ≦ Resource Potential=Sustainable Level
Year

Reservoir Engineering in Geothermal Fields. Figure 25
Forecast simulation to estimate sustainability of the resource exploitation

that predicts the decline of power output over time-assuming several scenarios of the development is commonly used to evaluate resource potential.

For example, in the case of the Hatchobaru geothermal power plant (110 MW), three scenarios were considered in which a total of 17, 26, and 31 makeup production wells were assumed to produce from the productive faults in the field. Both the number of wells and the spacing between them (100–200 m) were specified regardless of the actual number of wells that could be drilled from the existing drilling pads. These scenarios showed that the simulated total power output temporally reached 190–265 MW, depending on the number of assumed makeup wells; however, in all three scenarios, a power decline was experienced until it stabilized at around 120 MW (Fig. 26). The prediction period is 9 years, which may not be sufficient to determine the sustainable potential. Since a smaller number of makeup wells than assumed in this case study are required to maintain around 120 MW from beginning of the calculation, it will be easy to sustain around 120 MW over 9 years. Therefore, it is suggested that the sustainable power potential of the Hatchobaru reservoirs is approximately 120 MW. It will also be possible to maintain the rated power output of 110 MW if the targets for production makeup wells were selected in such a way as not to overproduce the faults beyond their sustainable power potential [2, 4, 11].



Reservoir Engineering in Geothermal Fields. Figure 26
Prediction of sustainable power output of the Hatchobaru reservoirs [2, 4]

Future Directions

Reservoir simulation will become a common technology to evaluate the resource potential and optimize the development scenario in the geothermal fields in the world [12–14]. In accordance with significant increase of computer capability, the reservoir simulator has been also improving and increasing its capability to perform more accurate reservoir simulation [15]. Three kinds of technical improvement and development have been currently endeavored. One is to enhance the simulation capabilities to calculate the transportation of a number of fluids; those are, for example, not only pure water but also brine, noncondensable gas, chemical species, volatile organic, and so on [16, 17]. In addition, the module of calculating chemical reactions between rocks and fluids has been developed [18–21]. Another one is to facilitate an inverse modeling application that enables to do an automatic history matching and parameter estimation based on data obtained during testing and exploitation of geothermal fields [22, 23]. The other one is a speedup of calculation for the entire reservoir simulation by adopting a parallel computing technology that enables multicomputers to be used simultaneously. Since the source program of the reservoir simulator includes subroutines that can be calculated in parallel, it is possible to reduce the calculation time by modifying the source program of those routines so that multicomputers can share the corresponding calculations. The above-mentioned capability improvement and expansion for the reservoir simulation will contribute to accelerate geothermal power development and reduce the entire costs, because accuracy of resource potential evaluation will increase, and therefore, the output of geothermal power plant will be successfully maintained.

Bibliography

Primary Literature

1. Axelsson G, Björnsson G, Montalvo F (2005) Quantitative interpretation of tracer test data. In: Proceedings of the World Geothermal Congress 2005, Antalya
2. Tokita H, Haruguchi K, Kamenosono H (2000) Maintaining the rated power output of the Hatchobaru geothermal field through an integrated reservoir management. In: Proceedings of World Geothermal Congress 2000, Kyushu – Tohoku, pp 2263–2268
3. Lima E, Tokita H, Tsukamoto S (2004) Operation of the Hatchobaru geothermal field based upon a coupled reservoir-well and piping network numerical simulator. In: Proceedings of the KenGen geothermal conference, Nairobi Kenya, CD-ROM
4. Tokita H, Lima E, Itoi R, Akiyoshi M, Senjyu T (2006) Application of coupled numerical reservoir simulation to design a sustainable exploitation of the Hatchobaru geothermal field. In: Proceedings of renewable energy 2006, Makuhari Messe, CD-ROM
5. Aunzo ZP, Bjornsson G, Bodvarsson GS (1991) Wellbore models GWELL, GWNACL, and HOLA user's guide. Lawrence Berkeley laboratory Report LBL-31428, DE92 009494, Oct 1991
6. Gunn C, Freeston D (1991) An integrated steady-state wellbore simulation and analysis package. In: Proceedings of the 13th New Zealand geothermal workshop 1991, Auckland, pp 161–166
7. Itoi R, Kakihara Y, Fukuda M, Koga A (1988) Numerical simulation of well characteristics coupled with steady radial flow in a geothermal reservoir. In: International symposium on geothermal energy, exploration and development of geothermal resources, Kumamoto and Beppu, Japan, pp 201–204
8. Takahashi M (1988) A wellbore flow model in the presence of CO_2 gas. In: Proceedings of the 13th workshop on geothermal reservoir engineering. Stanford University, Stanford, pp 151–157
9. Tokita H, Itoi R (2004) Development of MULFEWS a multi-feed wellbore simulator. In: Proceedings of 29th workshop on geothermal reservoir engineering, Stanford University, California, CD-ROM
10. Tokita H, Yahara T, Kitakoga I (1995) Cooling effect and fluid behavior due to reinjected hot water in the Hatchobaru geothermal field, Japan. In: Proceedings of the World Geothermal Congress 1995, vol 3, Florence, pp 1869–1874
11. Yahara T, Tokita H (2010) Sustainability of the Hatchobaru geothermal field, Japan. Geothermics 39:382–390
12. Pruess K (1991) TOUGH2 – A general-purpose numerical simulator for multiphase fluid and heat flow. Report LBL-29400, UC-251, pp 1–102
13. Pruess K, Oldenburg C, Moridis G (1999) TOUGH2 user's guide, version 2.0. Lawrence Berkeley National Laboratory, Berkeley, LBNL-43134
14. Prichett JW (1995) STAR: a geothermal reservoir simulation system. In: Proceedings of the World geothermal congress 1995, Florence, Italy, pp 2959–2963
15. Antunez E, Morids G, Pruess K (1995) Large-scale three-dimensional geothermal reservoir simulation on small computer system. In: World Geothermal Congress 1995, vol 4, Florence, pp 2977–2980
16. Battistelli A, Calore C, Pruess K (1997) The simulator TOUGH2/EWASG for modeling geothermal reservoirs with brines and non-condensable gas. Geothermics 26(4):437–464
17. Pruess K, Finsterle S, Moridis G, Oldenburg C, Wu Y (1997) General-purpose reservoir simulators: the TOUGH2 family. Geotherm Resour Council Bull 26:53–57
18. Kim J, Schwartz FW, Shi J, Xu T (2003) Modeling the coupling between flow and transport developed by chemical reactions and density differences using TOUGHREACT. In: Proceedings of the TOUGH symposium 2003, Berkeley
19. Sato T, Ohsato K, Shiga T, Sato M, White SP, Kissling WM (2003) A study of reservoir estimation for a deep-seated geothermal reservoir using TOUGH2 and CHEMITOUGH2. In: Proceedings of the TOUGH symposium 2003, Berkeley
20. Xu T, Sonnenthal E, Spycher N, Pruess K (2003) TOUGHREACT: A new code of the TOUGH family for non-isothermal multiphase reactive geothermal transport in variable saturated geologic media. Lawrence Berkeley National Laboratory, Berkeley, LBNL-52342

21. Xu T, Sonnenthal E, Spycher N, Pruess K (2003) Using toughreact to model reactive fluid flow and geochemical transport in hydrothermal system. In: GRC 2003 Annual meeting, Morelia
22. Finsterle S, Pruess K, Bullivant DP, O'Sullivan MJ (1997) Application of inverse modeling to geothermal reservoir simulation. In: Proceedings of the 22nd workshop on geothermal reservoir engineering, Stanford University, Stanford
23. Finsterle S (2007) iTOUGH2 user's guide. Lawrence Berkeley National Laboratory, Berkeley, CA94720

Books and Reviews

Grant MA, Bixley PF (2011) Geothermal reservoir engineering, 2nd edn. Academic, Burlington, 349 pp. ISBN 978-0-12-383880-3

Resource Repletion, Role of Buildings

Katja Hansen[1,2], Michael Braungart[1,2],
Douglas Mulhall[1,2]
[1]EPEA Internationale Umweltforschung GmbH,
Hamburg, Germany
[2]Academic Chair Cradle to Cradle for Innovation and
Quality, Rotterdam School of Management,
Erasmus University, Rotterdam, The Netherlands

Article Outline

Glossary
Definition of the Subject
Introduction
Unintended Consequences of Traditional Sustainability
Materials Repletion for an Abundant Healthy Footprint
Application Tools for Implementing Beneficial Resource Repletion
Redefining Recycling
Introducing Nutrient Certificates
Conclusions
Acknowledgments
Bibliography

Glossary

Improving the approach to materials and products sometimes requires revising traditional terminology. In the approach described here, usage of certain terms differs from traditional definitions, to account for innovative features of materials and products.

Biobased vs. biodegradable Many biobased products such as, for example, biopolymers are not necessarily safely biodegradable because they contain additives such as heavy metals or are combined with nonbiodegradable materials. As well, petroleum-based products that are not biobased can be biodegradable. So it is important to distinguish these features to develop an effective defined-use pathway for materials. Especially, it is important to evaluate biobased and biodegradable in the context of the intended use of the material, e.g., if it is intended for a biosphere or technosphere pathway. For example, many materials designed for single use before disposal in a biosphere pathway and defined as biodegradable, such as cups, do not biodegrade in the processing time frame used in an industrial composting facility and, as a result, end up being incompletely decomposed and incinerated, or degrade the quality of compost. Because of this, the definition of "biodegradable" includes that the material is shown to degrade completely in an industrial composting facility within a prescribed time frame.

Counter-footprint Calculation showing activities that can be used to counterbalance a negative "environmental footprint." Example, producing renewable energy instead of just consuming energy. Counter-footprinting is still at an early stage and often, for example, does not calculate defined material content, defined-use pathways, or beneficial functions of materials such as, for example, cleaning the air. For example, Coto-Millan et al. [1] list construction materials as resource consumption, but not as a material resource on the counter-footprint side of the equation. The part of land "consumed" for structures is regarded only on the negative side of the footprint equation, rather than as a productive contributor to the ecology. In general, when materials are used for constructing a building, their impacts are frequently still considered only on the negative side of the environmental footprint and no longer considered as beneficial resources. See also "offset."

Cradle to Cradle® An innovation platform to improve the beneficial qualities of products and services in biosphere and technosphere metabolisms as a step beyond the traditional sustainability approach of reducing negative impacts. The term *Cradle to Cradle®* is a registered mark for quality assurance purposes, similar to how the broadly accepted International Standards Organization governs use of its marks and standards. However, the philosophy, principles, and many application tools of the Cradle to Cradle® approach are widely published. The founders of the C2C approach encourage governments, companies, and NGOs to use the philosophy and principles. The right to use the Cradle to Cradle Design Protocol® for certification is assigned to an independent nonprofit organization, and certification criteria are also broadly published.

Defined use Materials and products that are designed according to their intended use in biosphere or technosphere metabolisms.

Depletion Loss of nonrenewable resources and destruction of renewable resources.

Ecological footprint Usually a calculation of negative environmental impacts of human activity. Many definitions are used, but an example in relation to the built environment is "Corporate ecological footprint is defined as the environmental impact (in hectares) of any organisation, caused by: (a) the purchase of any kind of product and service clearly reflected in their financial accounts; (b) the sale of products deriving from the primary production of food and other forestry or biotic resources, or in other words when vegetables, fruit and meat enter the market chain for the first time; (c) occupation of space; and (d) generation of waste clearly reflected in their environmental report. Moreover, this impact measured in hectares can be transformed to obtain a result in tons of CO_2 emitted (the carbon footprint)..." [1]. See also Counter-footprint and Offset.

Intelligent materials pooling (IMP) Sharing of defined material streams among partners to achieve economy of scale and accelerate the use of C2C-defined materials.

Materials bank Database-supported pool of defined materials.

Materials security Security of supply for strategically important materials such as rare metals or phosphate.

Nutrient certificates Set of data describing defined characteristics of materials in products that give them value for recovery and reuse. Nutrient Certificates are a marketplace mechanism to encourage product designs, material recovery systems, and chain of possession partnerships that improve the quality, value, and security of supply for materials so they can be reused in continuous loops or closed loops or beneficially returned to biological systems. This is done by adding a new value dimension to materials quality. This new dimension is based on the suitability of materials for recovery and reuse as resources in other products and processes.

Offset Assessment of activities that compensate for negative environmental impacts. As opposed to counter-footprints, offsets are often used to describe remotely located activities, such as growing trees in another location to replace trees lost due to development. However, counter-footprint and offsets can also overlap.

Recycled vs. recyclable Products can be beneficial if they have defined recyclable content regardless if it is recycled or not. Defined recyclable content is an enabler for recycled content. If virgin content is not recyclable then it will pollute recycling streams, so recyclable is just as important as being recycled. Recycled content that is also recyclable at a similar level of quality is the end goal of product design for Nutrient Certificates.

Recycling There are many definitions of recycling, but for these purposes, recycling is defined as recovering and reusing materials at a similar level of quality by defining their content, as compared to "downcycling" where materials are recovered and reused at a lower quality level. For example, the term "recycling" is often applied to materials such as paper, but in reality, paper is almost always downcycled due to shortening of its fibers. Many current definitions of recycled content do not define what is in the material, with the result that it is not possible to recycle the materials at a similar level of quality. The important distinguishing factor is "defined" content, which can be indicated as defined to 100 ppm.

Repletion Replenishing the supply of biosphere and technosphere materials for use in products and processes.

Scarcity Geographically, politically, or commercially limited supply of strategic materials.

Upcycling Improving the existing quality of a material for its next reuse. A material can be defined as upcycled under various conditions:

(1) When its current downcycling is improved so the material is recycled at a similar level of quality instead of lower level. For example, high-grade steel is separated from motors containing copper contaminants so the steel can be resmelted at the same level instead of downcycled

(2) When a degraded material is repaired for effective reuse, e.g., an additive is added to a plastic to repair its damaged molecular strings so the material can be reused for a high quality purpose

Definition of the Subject

Raw materials scarcity, rising raw materials extraction costs, and biodiversity loss are apparent globally. Recycling of materials is cited as one solution to those problems. However, maintaining the consistent quality of materials is excluded from most traditional sustainability assessments, and current regimes of carbon, emissions, and energy trading are not well designed to account for the quality or value of materials, or the processes for achieving materials recovery and reuse.

The building industry is a large consumer of scarce resources, and because of this, it is regarded as a leading cause of resource depletion. However, at the same time, materials contained in and moving through buildings have been extensively evaluated for their recovery potential [2], and as a result, could be used in a new model where buildings are resource repleters instead of depleters. Materials repletion is a value-based business model that defines new dimensions of quality to generate quantifiable benefits for builders, suppliers, building occupants, operators, and owners. For this model, the inherent conservatism of the industry could work in favor of a new approach due to the emphasis on value, reliability, and documentation. This would introduce a new type of beneficial agenda into the supply chain of the construction, operations, maintenance, decommissioning, and recycling industries.

To support this beneficial paradigm, a framework for resource repletion and security is described where materials are defined according to qualities that enhance their bankable and tradable value. The concept of *Nutrient Certificates* is introduced as a counterpart to emissions certificates to account for the value of defined high-quality material flows. The focus is distinct from mechanisms such as environmental product declarations, material safety data sheets (MSDS), or emissions trading. For this approach, new criteria are introduced for materials and for "recycled content." Financial enabling innovations are described, and early adopters are identified among governments and companies, including participants in Cradle to Cradle® networks.

Introduction

> ▶ Our economy will run out of materials before it runs out of energy.
>
> (Michael Braungart 2009)

> ▶ China. . .has now quietly halted some shipments of (rare earth metals) to the United States and Europe. . . industry and government are joining forces by appealing to the European Commission and the World Trade Organization to intervene. . .
>
> (*New York Times* October 19, 2010 [3, 4])

The globalized depletion of material resources is well documented quantitatively and qualitatively [5]. Skeptics argue that such claims are exaggerated [6] and that scarcity has a positive side effect of driving innovation. Despite those arguments, scarcity is still disrupting economies. This scarcity is generated more by geopolitical distribution inequity than geological availability. For example, Europe depends on just two countries, China and Morocco, to support most of its agricultural system with the fertilizer phosphate. In 2008, China puts export tariffs on its phosphate supplies, disrupting phosphate pricing [7]. China also has a near-monopoly on producing certain rare metals, not because of geographic distribution but because other countries let their rare metal processing capacities dwindle [8, 9] via outsourcing. The city of Barcelona once had to import water in ships due to a water shortage, resulting in prioritization conflicts with other regions of Spain [10]. As well, Central and Southern Europe suffered when Russia interrupted gas supplies.

In addition to distribution disruptions, the costs of extracting material resources are accelerating as supplies become more remote and difficult to access. Biodiversity losses are accelerating as species habitats are put under greater pressure by expanding incursion for urbanization and resource extraction [11].

Together, those factors pose economic security risks to regions such as Europe, USA, and Japan, along with suppliers such as China who come under pressure to solve shortages by exporting their own limited supplies. Materials security is taking the stage alongside energy security as an economic and military consideration. This is generating an economic, social, and ecological imperative for materials repletion instead of depletion.

A first step to establishing a materials repletion paradigm is to recognize the unintended consequences of traditional approaches to "sustainability," which pose barriers to repletion. The step after that is to describe solutions for those unintended consequences.

Unintended Consequences of Traditional Sustainability

The Unintended Consequences of Eco-Efficiency

The traditional approach to materials sustainability in buildings is exemplified by the draft topical outline provided to the authors for this article:

- Green sustainable building is the practice of increasing the efficiency with which buildings and their sites use and harvest energy, water, and material, while at the same time reducing impact on the local and global natural environment. [12]

An unintended consequence of this approach has been that greater efficiency often has the opposite effect by accelerating instead of reducing impacts. This is self-evident, for example, with accelerated material and energy throughputs for personal electronic devices. Here, efficiency improvements in the use of materials and energy have been exceptional over past decades. Miniaturization resulted in devices that require only a fraction of the energy and the materials compared to earlier devices to perform the same function, while at the same time adding functional features. However, this resulted in price drops that made the technology accessible to billions of customers, generating exponential increases in materials and energy throughputs due to collective demand [13]. By increasing the efficiency of individual devices, the industry accelerated collective materials and energy throughputs, including flows of materials for products to waste from product disposal.

The same occurred with buildings, as efficiencies have improved and economic activity expands globally, real growth in the collective "ecological footprint" of buildings (see Glossary) has accelerated in developing economies of Eurasia, Africa, the Middle East, and South America, while still expanding in developed economies of Europe, North America, and elsewhere. The results are marginal reductions in the footprints of individual buildings negated by the collective expansion of land use, materials, and energy. In many cases, individual building footprints are increasing as well to house diverse technologies and the lifestyles they suggest. Electronic technologies and networking services in particular are growing rapidly, resulting in increasing demands for space, materials and energy.

This collective expansion of impacts due to economic and technological development is also described in recent studies finding that wealth, not poverty, is the main driver of environmental impacts [14]. As collective wealth improves globally, consumption of energy and resources is accelerating.

Due to the phenomenon of efficiency accelerating throughputs, the "reduction" approach to resources on its own is unlikely to significantly mitigate environmental impacts in buildings or of products that move through them, and instead is likely to accelerate those impacts.

Disregarding Beneficial Footprints

- Minimization vs. Benefits. Green assessment methods focus more on minimizing use of materials and energy instead of focusing on the intended benefits of materials and energy in a system. As a result, the value of some beneficial aspects is underemphasized in green building standards. By contrast, natural processes often maximize their own footprint, and an example of this is the largest living thing on Earth, the giant sequoia. See following sections for example.
- Measurement Criteria. To accurately assess the benefits of some materials, it is important to

evaluate factors such as their "defined use" in ecosystems, the active benefits they generate, and the time frame and rate of reuse, as discussed under Section III. Current measurement systems often disregard those.

- Assessing Impacts vs. Calculating Resources. Most green rating systems for buildings do not award many points for designing materials as resources for the future. Instead, those materials are often considered as part of a negative "consumption" footprint. Until now, this has been justifiable because many building products today are not designed with next use in mind, so they are not considered beneficial resources. But the unintended consequence is that when green rating systems do not encourage the transformation from minimizing impacts to maximizing future resources, they become part of the problem instead of the solution. An example of this can be found with "recycled content."

Unintended Consequences of "Recycled Content"

Part of the approach of green rating systems to materials depletion includes "recycled content" and "reuse of materials." However, those rating systems do not distinguish, for example, materials downcycled from high to low quality from materials with recycled or "virgin" content that is defined, easily separated, and recyclable at a similar level of quality. For example, LEED 2009 for existing buildings does not require "postconsumer recycled content" to be defined or to be recyclable at a similar level of quality [15]. Also BREEAM 2010 standards for real estate specify materials with a "high recycled content" and "reuse of materials," and although these also specify "nontoxic" materials, the standards do not specify defined or recyclable at a similar level of quality [16].

As a result, companies that invest in important quality aspects such as design for materials recovery are sometimes penalized for focusing on those investments. For example, well-defined virgin material for ongoing use would not receive credit, whereas undefined downcycled material would. Nor are actively beneficial functions of materials considered. So, a building containing materials defined to 100 ppm and consisting of surfaces that clean the air would not qualify under many green building evaluation systems because those materials do not contain "recycled" content. The defined content for ongoing use and the functional qualities of materials might be more beneficial for the building than undefined recycled content, but those benefits are underemphasized in Green building scoring systems and regulations.

This is especially problematic if contract tendering documents require contractors to maximize the amount of materials recycled or reused from an existing building into new construction or renovation. By mandating maximal on-site reuse or recycling, contracts can inadvertently penalize contractors who use well-defined materials wherein the next intended use as agricultural or industrial resource is known. Tenders can also discourage other innovative off-site reuse or recycling of materials that might be more suitable, including the reuse of those materials unsuitable for human exposure in a building.

> This is not to say recycled content is a bad requirement, but rather that the traditional definition could be improved to reward other properties and promote the use of defined and truly recyclable materials.

Likewise, material safety data sheets (MSDS) for products contain a list of aspects defined as hazardous and how to handle those, but are not designed to show the beneficial qualities or whether a product is sufficiently "defined" to be recycled at a similar level of quality. Nor are MSDS intended to evaluate the suitability of materials content for recycling after their first installation. Indeed, a material might be contaminated with an additive that makes it unsuitable for recycling, but this would not be described in the MSDS. Nor do MSDS keep up with science effectively. They are based on regulatory definitions that usually take many years to recognize hazards in products, long after the hazard has been established through scientific analysis. The effect of this is to increase contingent liabilities for companies because something that might not be listed on an MSDS today might suddenly appear when legislation changes. So MSDS provide no early warning system to prepare companies and owners for new findings.

Consequences of Traditional Assessment Boundaries

Beyond the confines of product selection, very little attention is paid to what happens to the products and materials that flow through buildings. Over the lifetime of a building, flow through might equal or exceed the mass of its construction materials. Often, no score is assigned to the quality of material in office equipment, furniture, and finishes that flows through the building, including food, paper products, and other biogenic materials, as illustrated in Fig. 1 in a following section.

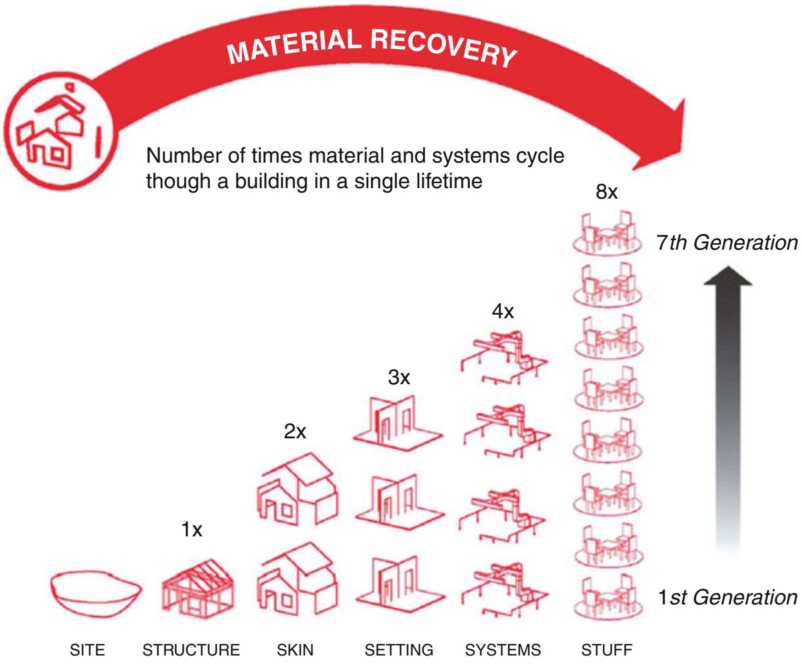
The emphasis of assessment language on being "less bad" is endemic throughout the literature. For example, the terms "minimize" and "reduce" are used throughout LEED design innovation credit documents [17]. Terms such as "offset" are used to signify compensation for negative impacts, and reference is also made to habitat conservation, but those approaches are rarely used in reference to materials quality. Nor are they intended to transform buildings into net positive generators of biodiversity, energy, or defined materials. Instead they are largely directed at reducing or compensating for the negative impacts of a building.

The effect of this language is that most attention and funding is spent on minimizing the negative impacts of building materials, and less attention is paid to evaluating or improving the benefits of material quality.

Such examples are not intended to discredit green building ratings as a mechanism, but instead, to show that limiting boundaries and emphasis can skew the focus of investment and limit the capacity to be truly beneficial. When boundaries are too limited, innovation is limited.

Unintended Consequences of Biomass Energy

The value of biological materials in buildings is usually considered in terms of how much wood and other



Resource Repletion, Role of Buildings. Figure 1
Number of times materials and systems cycle through a building in a single lifetime. It can be seen here that buildings are effective mechanisms for tracking materials with relatively short turnover periods as well as those with extended turnover periods, and that the shorter-term innovation potential is with products that have shorter generations for optimization (Source: Redrawn from a William Mc Donough & Partners drawing adapted from Brand 1994 [44])

agricultural products are used during construction. However, one of the largest components of biomass use for buildings, and one of the fastest growing in terms of investment, is biomass for energy [18]. It is noteworthy that technology has progressed over the centuries toward increasingly intensive embodied-energy fuel sources, from biomass to coal to oil to uranium, but now seems to be regressing to biomass. The widespread burning of virgin wood and other biomass is now sanctioned as "sustainable" and is given government incentives. However, the impacts of that burning whether in Africa or Europe are quickly becoming apparent in resource depletion. For example, a FAO study estimates that despite uncertainties over data, an approximately 400 million cubic meter shortfall by 2020 can be projected for wood availability in Europe [19]. This forecast was already manifested in 2010 when some biomass incineration facilities had to start limiting operations due to supply shortages. Europe would have to cut down many of its forests and use much of its farmland to meet this shortfall, so instead, it is importing wood from the Americas, Russia, South-East Asia, and Africa to meet demand [20]. Moreover, the fertilizer requirement for growing such large amounts of biomass is usually not considered in forecasting e.g, the capacities of soils to continually produce biomass that is not returned to the soil. The costs and materials flows involved in replacing humus, NPK, and trace elements are often absent from biomass supply chain estimates.

> This is not to say biomass energy is inherently impractical for meeting a minor portion of energy demand, but rather that current strategies are already leading to shortages and have not adequately considered nutrient flows required to meet production requirements. Nor do they consider the unintended consequences for the wood construction, pulp and paper, and furniture manufacturing industries that find themselves adversely affected by subsidized biomass burning [21].

Confusion Over Biobased Materials

Large investments are being directed into biobased materials, but at the same time, there is much confusion in the marketplace over what biobased materials are and under what conditions they are eco-friendly. "Biobased" is often confused with "biodegradable." Biobased materials are often not safely biodegradable due to the additives they contain. Nor are biobased materials necessarily eco-friendly. They can compete with food supplies and lead to topsoil depletion. To solve this, a systematic implementation regime is required that can define conditions under which biobased materials are beneficial along their intended pathway of production, use, and disposal.

Impracticality of "Local Materials" Criteria

In a globalized economy, sustainability criteria that reward local sourcing of materials do not adequately address regional materials imbalances. For example:

- Rare-earth metals required for industry are found in just a few countries.
- "Fair trade" regimes often involve transport of goods long distances to locations where similar resources are impractical to produce, e.g., coffee grown in Africa would require vast greenhouse complexes in Alaska.
- Innovative materials such as, e.g., Nanogels® that improve the performance of buildings are only manufactured in a few locations. Green building assessment systems such as BREEAM give preference to "local materials" [16]. Under those rules, innovative materials could be penalized.

While sourcing local materials, components, and labor can sometimes ensure reduced transportation costs, enhanced employment, and industrial durability in a region, and may source more climatically appropriate materials, those goals need to be modified to recognize where local sourcing is not the most practical alternative.

Deficiencies of Emissions Trading Schemes

Because emissions trading schemes and habitat offset schemes are usually characterized in terms of energy and wildlife, these are usually not referred to in relation to materials quality. However, those schemes are largely materials dependent. For example, fuels are materials. Energy systems in and for buildings are made of materials. Substances that make up habitats are materials. Greenhouse gases are made of chemicals that can be reused. Those examples suggest that emissions trading schemes are as much materials depletion and repletion questions as they are climate change or energy questions.

Emissions trading schemes are inadequately designed to promote conditions necessary for regenerating materials resources. For example, there are few if any emission credits for directly regenerating our primary carbon-sequestering systems, such as topsoil which is the upper 1–2 m of soil required by most plants to grow. Soil organic carbon is the biggest carbon pool of the planet after the oceans and far greater than what is found in the atmosphere [22]. Soil is a forgotten climate solution. Aside from the oceans, top soil is the leading repository of carbon in the biosphere. Much of the productive topsoil globally has been lost in the past century due to industrialized farming, soil compaction, erosion, and urbanization [23]. This loss has been a leading contributor to carbon release into the atmosphere. Conversely, soil conservation is shown to sequester atmospheric carbon [24]. Emissions trading schemes can be used to support soil conservation, which is distinct from soil manufacturing, but they do not account for the nitrogen, phosphorous, potassium (NPK) fertilizer cycle in topsoil, or soil quality degradation from rock phosphate fertilizer that is contaminated with uranium mined with the rock phosphate [25].

Topsoil is often not considered as a significant factor in buildings, but actually, buildings and topsoil have substantial interactions. Buildings usually occupy space where topsoil used to be. Runoff from buildings has substantial impacts on topsoil. Landscaping for buildings makes extensive use of topsoil. Products used in buildings are often made from biomass grown in soil. So it can be said that buildings rely heavily on topsoil. Because of this, converting buildings from materials depleters to materials repleters also involves a basic revision in the approach to soil.

Due to those various factors, emissions trading schemes are ill-equipped to support or quantify materials flows that are essential for the continued development of our industrial society. However, it is not the intention here to describe all the pros and cons of such trading schemes, as these are described elsewhere [26].

Instead, the main purpose here is to point out that those systems often do not recognize the role of materials. They often do not quantify the contribution of materials to carbon sequestration, biodiversity, nutrient recycling, and other beneficial aspects. They do not distinguish between high quality and low quality recycling.

Materials Repletion for an Abundant Healthy Footprint

How might it be possible to solve the unintended consequences of traditional sustainability approaches? Practical lessons can be taken from natural systems.

Big Footprints Can Be Beneficial

One example of a big, healthy footprint is the largest living thing on Earth, the giant sequoia [27]. Like many buildings, it reaches heights of more than 90 m, weighs thousands of tons, and uses thousands of liters of water and large amounts of energy daily. As the fastest growing tree in the world, it consumes large amounts of energy to pump nutrients and relies on the massive transport of water from hundreds of kilometers away to provide the marine layer that feeds the forest canopy. Its outer skin is far from minimal; it is thicker than the outer skin of many buildings and can be more than a meter thick. Its root system extends farther underground and laterally than the foundations of many buildings. The total throughput of energy and materials of a forest containing giant sequoia and similar giant tree species can rival that of small human settlements. In total, the giant sequoia has a giant footprint.

The sequoia also depends on destructive processes. It depends for its early existence on fires that pollute the environment and kill other trees, clearing the forest floor so sequoia species can compete and get established. To protect itself, it uses the toxic material tannin which can be harmful to insects and other wildlife. At the same time though, the sequoia has a maximal beneficial footprint, which can be described as an "offset" or "counter-footprint" [28] (see Glossary) that maximizes the positive use and reuse of resources. It provides a habitat for hundreds of species, generates oxygen, beneficially uses CO_2, filters the air, is an exceptionally long-term carbon sink, and sheds biomass that is converted into soil. Its cellular capacities allow it to use large amounts of water while returning it to the environment in a state beneficial for other biosystems. Its immune system resists disease, letting it live more than 3,000 years. The species is so genetically successful, it has outlived the dinosaurs.

Maximizing Benefits is More than Just Semantics

While the distinction between "maximizing benefits" and "minimizing impacts" might seem like semantics, it extends far beyond terminology. Natural design principles are increasingly recognized by architects and designers and applied to building and product designs through techniques such as biomimicry and calculation of offsets or counter-footprints. When innovation is directed at maximizing benefits instead of minimizing damage, different outcomes occur. For example, instead of minimizing the amounts of rare metals in devices to a level where it might not be economic to recover them, the designer focuses on design for disassembly to recover those rare metals so they can be used again. This approach might in some cases increase the amounts of rare metals used, but also enables their recovery. In this way the materials can be used as resources for other industrial processes.

This different approach can be seen in the Cradle to Cradle® approach to resource reutilization.

Cradle to Cradle® for Achieving Resource Repletion and Beneficial Footprints

The Cradle to Cradle® (C2C) [29] approach emphasizes those beneficial factors. The "Cradle to Cradle" Design Protocol® has taken an approach that generates benefits for stakeholders by going beyond the "cradle to grave" and beyond traditional interpretations of "environment."

Cradle to Cradle® is a paradigm-changing, quality-enhancing innovation platform developed in the 1990s by Michael Braungart, William McDonough, and others based on research at the Environmental Protection Encouragement Agency [30] in Hamburg Germany, for designing beneficial economic, social, and environmental features into products, processes, and systems. Cradle to Cradle® is primarily an entrepreneurial and innovation approach that starts by determining the intended benefits of a product or service instead of focusing on minimizing negative environmental impacts.

C2C philosophy, principles, and many of its application tools are broadly published, and the philosophy and principles are available for anybody to use with attribution.

The C2C Design Protocol described here, as well as the *Cradle to Cradle® Criteria for the Built Environment* are further described in other publications and are only partially excerpted here [31].

To enhance quality and add value for stakeholders, C2C promotes innovation partnerships along the entire chain of a product, including manufacturing, distribution, use, disassembly, recovery, and reuse.

By characterizing hundreds of products and thousands of materials for their human and environmental health attributes, as well as defining systems to safely and fully cycle materials into new products, C2C has already provided a practical yet inspirational scientific and business model for improving quality.

This innovation and value model makes C2C potentially attractive to planners, builders, and manufacturers for integration into products, processes, buildings, materials recovery systems, and purchasing.

Extensive books, cover stories, and documentary films have been published and broadcasted about C2C since the 1990s. The book *Cradle to Cradle* [32] is well known and translated into at least a dozen languages. However, many planners are not yet familiar with how to integrate into the built environment C2C features such as beneficial materials. There is a tendency when encountering well-known phrases such as "safe materials" and "species diversity," to respond with "yes we do that already." But most buildings and area plans do not incorporate defined-use pathways for materials. Methods are still not well established for designing sites so they contain "defined" materials, or are species-positive.

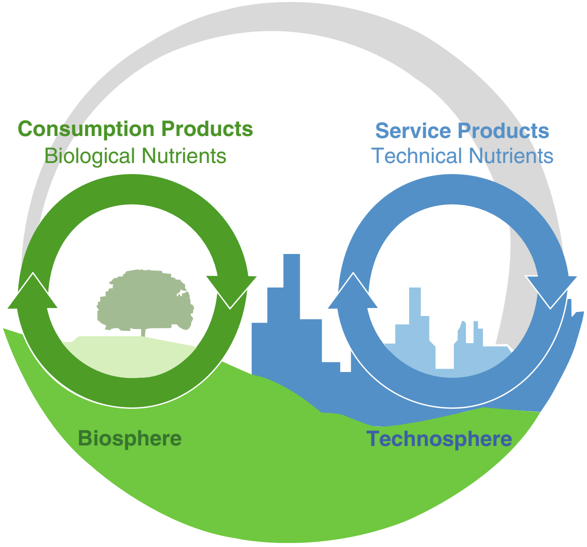
Various guidelines for C2C in the built environment were integral to published declarations such as the Hannover principles [33] and more recently in The Netherlands, the Almere principles [34].

Those extensive documents are only effective if they can be translated into measurable results. The first step is to understand the overall C2C framework, then study and implement the three defining Cradle to Cradle® principles encompassed in that framework.

C2C Framework
Cradle to Cradle® can be divided into these categories that together make up the C2C Framework;
(1) Philosophy e.g., a quality-based innovation platform for benefitting the economy, ecology and social equity.
(2) Principles that are translated into measurable criteria.
(3) Application tools.

C2C Principles
Waste = Food. Everything is a Nutrient for Something Else.
Use Current Solar Income. Energy that can be Renewed as it is Used.
Celebrate Diversity. Species, Cultural, and Innovation Diversity.

These principles define and support two types of metabolism for every product and process; biosphere metabolisms for products designed to support biological processes, and technosphere metabolisms for products designed to provide a technical service and whose materials are continuously recycled. See Fig. 2.



Resource Repletion, Role of Buildings. Figure 2
Cradle to Cradle® metabolisms and product types

Using the C2C principles as a basis, a Cradle to Cradle® building can be defined this way:

A Cradle to Cradle® building contains defined elements that add value and celebrate innovation and enjoyment by measurably enhancing the quality of materials, biodiversity, air, and water, using current solar income, being deconstructable and recyclable, and performing diverse practical and life-enhancing functions for its stakeholders.

The definition also applies to materials and products such as furnishings, cleaning materials, and office equipment that move through buildings because, often, things that move through buildings have equal or greater impacts on the economy, environment, and users than the structures themselves. This “moveables” factor is often underemphasized in various “green” guidelines for buildings, but it is a central feature of the Cradle to Cradle® approach.

Ten Criteria for Applying and Quantifying C2C Principles in Buildings

Broad principles are a prerequisite for eco-effectiveness. However, measurable criteria are also important for practical implementation. The following ten criteria for applying C2C principles have been organized to facilitate implementation.

Through these criteria, materials in buildings can be defined to constitute a new value chain (1) themselves, (2) through the services they provide in a building, such as cleaning air. For example, defined materials often have a greater recovery value than undefined ones due to their enhanced qualities, but it can also result in cleaner air and water flowing through a building. The quality of air, water, biological nutrients, and biodiversity each constitute value streams that make buildings beneficial contributors. Integrating those aspects also enhances the overall value of a building because the whole is often greater than the sum of its parts. For example, a building whose materials clean the air and water reduces the costs of conventional purification systems.

The ten criteria constitute a guide for achieving those value streams.

State Your Intentions Design is the first signal of human intention. To achieve truly eco-effective solution sets, it is important to design so that; every material can be a nutrient for the next design, every element is generated within our solar income, and the design embraces human- and bio-diversity.

Examples of those design intentions: Do you want the building to contribute air and water that are cleaner than when they were taken from the outdoor environment? Do you want the building to be deconstructable? Do you want to demonstrate that the ingredients in building materials are defined and safe?

Define Materials and Their Intended Use Pathways

(a) Use materials whose quality and contents are measurably defined in technical or biological pathways from manufacturing through use and recovery.
(b) Use materials whose impacts are measurably beneficial for human health and the environment.

For example, a defined product would be a chair whose component parts come from known renewable or recycled materials and energy sources, whose composition is known to 100 parts per million, whose materials are safe for contact with human skin and lungs and can be disassembled into materials that each can be recycled for use in other products or decomposed as beneficial nutrients for biological systems. A "beneficial" ingredient would be an ingredient added to coatings that allows them to actively clean the air.

Integrate Biological Nutrients Measurably recycle biological nutrients and water by integrating biomass production into buildings, landscaping, and spatial plans to generate more biomass, soil, and clean water than before development of the site.

For example, biological nutrients from gray water, biodigestion, and interior and exterior landscaping can be recaptured. Air-cleaning vegetative walls can be designed to metabolize pollutants, and "green roofs" can be designed to retain moisture, capture CO_2, metabolize particulates, and provide oxygen. "Topsoil manufacturing" can be integrated into projects to use biodigestion and composting to produce humus and capture CO_2.

Enhance Air and Climate Quality

(a) Measurably make air quality healthier for biological metabolisms than before it enters a building and provide a comfortable climate for occupants.
(b) Measurably contribute to enhancing outdoor climate by contributing air that is healthier for biosphere metabolisms than before it enters a building and using climate change gases as resources through carbon management.

For example, (a) Air quality can be enhanced by integrating C2C materials across products such as exposed window frames, floors, wall materials, HVAC systems, wall and floor coverings, indoor plants and green walls, furnishings, office equipment, and mold inhibitors. (b) Active carbon management is achieved with vegetation and renewable energy. Climate change gases such as methane and CO_2 are resources that can be used to produce biomass, to be further discussed in later sections.

Enhance Water Quality Measurably improve water quality so the water is healthier for biological metabolisms than before it entered the building.

For example, water quality improvement can be achieved by integrating water recycling systems with nutrient recycling, rainfall capture and storage, indoor plants, and green walls.

Integrate Renewable Energy Integrate renewable energy (current solar, wind, and gravitational income) into buildings and area plans so that the building and site generate more energy than they use. Use exergy as a way to guide energy effectiveness.

For example, integrate "threshold" energy efficiency used in high-efficiency LED lighting with direct current from photovoltaic cells, daylight, ventilation, solar heating, and cooling as renewable energy sources.

Actively Support Biodiversity Integrate measurable species diversity so the area supports more diversity than before development.

For example, species diversity applies to plants, animals, and insects and is quantified by counting numbers and varieties supported by a building. The concept of "natural" or "native" species has to be evaluated in each case because in many regions, the natural environment has been transformed by humans, and returning it to an earlier "natural" state might be impractical.

Support Diversity with Innovation Pursue diversity innovations by focusing on special beneficial features of a building and integrating innovative components that are beneficial for the well-being of occupants and the environment.

For example, diversity gains can be quantified by the variety and prevalence of materials designed as nutrients in a building, the percentage of energy used that is truly renewable, and the amount of beneficial air, water, topsoil, and biodiversity contributed to the outside environment. "Buildings like trees" is a guiding C2C innovative approach. Some innovation can be achieved through biomimicry, e.g., coatings that metabolize pollutants, while others require a level of systems integration that rival the giant sequoia.

Add Value and Enhance Quality for Stakeholders In addition to generating value for the general environment and population, describe what the C2C features of a building do practically for the building owners, operators, and occupants.

For example, cleaner indoor air enhances productivity; recycling water reduces water fees; building integrated photovoltaics can be less expensive than other claddings while providing energy security in regions with irregular power supply; design for disassembly of HVAC systems supports inexpensive replacement during the life of the building; natural lighting cuts energy costs and can enhance human health.

Enhance quality of life for diverse stakeholders Enhance quality of life by designing C2C materials and integrated systems to support sociocultural richness.

For example, innovations can meet C2C principles and make areas safe for children, enhance accessibility, and provide ready access to outdoors and fresh air.

Application Tools for Implementing Beneficial Resource Repletion

Context for C2C Application Tools

C2C principles are primarily qualitative. The criteria related to those principles described earlier are designed to allow quantification of those qualities in buildings.

To implement the criteria and to solve many of the unintended consequences associated with conventional green building criteria, a number of application tools have been developed. These are described more completely in the publication *Cradle to Cradle® Criteria for the Built Environment* [31].

Those application tools are guided closely by C2C principles and criteria. This link distinguishes their use in C2C from how they might be used under conventional sustainability approaches.

For example, C2C approaches CO_2 and energy as *materials opportunities* rather than problems to be "capped and traded" or "captured and stored." Instead of focusing on cap and trade or carbon capture and storage (CCS), C2C application tools would focus on carbon capture and reuse (CCR).

A brief explanation of the C2C approach to energy and CO_2 is an appropriate example, to demonstrate how C2C application tools are applied differently from conventional sustainability approaches;

Cradle to Cradle® (C2C) energy is energy that is generated and applied effectively, using current solar or gravitational income, and material media that are defined as biological or technical nutrients. The definition is qualified and quantified by the following criteria together:

- *Energy sources.* Use current solar or gravitational income, or other defined C2C sources. Primary examples of current solar income use, conversion, and storage include natural light, solar thermal, photovoltaic, photochemical, wave and wind energy, thermal mass storage, and heat exchange. Secondary solar uses include, currently renewable biomass-derived energy from composting, biodigestion, thermolysis, hydrothermolysis, pyrolosis, gasification, and fuel cells. Gravitational income examples, kinetic energy from inertia or weight, e.g., descending waterways. Each of those energy sources is also evaluated under C2C according to the defined use of the material media.
- *Material media.* For generating, converting, and using energy, use materials that contain defined biological or technical nutrients at each stage, converted to energy in the final stage of the material cascade.

- *Energy effectiveness*. Generate and use energy in definably effective ways, using exergy as a way of measuring effectiveness.

From the C2C perspective, carbon dioxide and related gases are a resource. Surprisingly, many methods used to calculate "carbon footprint" of buildings do not include the beneficial use of carbon by, for example, vegetation. Although some carbon footprint methods are beginning to include "counter-footprint" calculations, the concept that buildings themselves could be beneficial users of carbon has not made its way into methodologies.

Primary Application Tools

Because materials and nutrients are the basis for energy generation as well as product manufacturing, the following sections describe three material-related application tools for achieving beneficial resource repletion in buildings including the products and energy devices that are used in buildings:

- Redefining recycling to include defined content and intended use.
- Introducing Nutrient Certificates to quantify the contents and benefits of materials.
- Quantifying and applying economic benefits from those improvements.

Redefining Recycling

The definition of "recycled content" could be improved by including these factors that describe the quality of a product:

Content

What is in the recycled content? For example, are all contents known, especially additives that give materials such as paper, plastic, and metals added functional qualities?

- The material description distinguishes "recycled" from "recyclable" content and "biobased" from "biodegradable" content (see Glossary).
- In the case of recyclable content, the material has infrastructures in place for recovering content.
- For biodegradable content, the material can decompose in available biodegradation facilities. For example, many biodegradable materials do not decompose fast enough in industrial composting facilities and are incinerated. To solve this, the material is defined for industrial composting.

Intended Use

What the material is intended to do and the pathway it is defined for. Intended use is described later in Cradle to Cradle® biosphere and technosphere metabolisms. For example:

- The material is used once then discarded into biological pathways, or is used repeatedly in technological pathways.
- The material can contain toxic materials that perform a function but only if they are safely locked into the material and can be recovered safely for reuse.
- The material performs a beneficial function such as cleaning the air or generating renewable energy.

Material Integrity

- The material can be recycled at a similar level of quality as distinguished from downcycled into lower quality products. If the quality cannot be maintained, e.g., with paper, can the material be downcycled in a controlled "cascade" so its uses are maximized? See *Nutrient Certificates* and *Material Cascades*, as well as Glossary for clarifications on recycling terminology.
- The materials are assembled into the product in such a way that their integrity can be recovered when the product is disassembled.
- If the material is made from recyclable virgin content, then it has a designated pathway for recycling at a similar level of quality.

Distinguishing between Renewable and Recoverable

Sometimes recycling information includes a statement that the material comes from "renewable" resources. However, it is more important whether and how the materials can be recovered and reused as nutrients. This avoids the "renewability" designation being undermined by increasing demand vs. supply of a given material. For example, some natural fibers that only grow in certain regions might be renewable the first year then unsustainable the next when billions of people use them in their products. Wood is renewable until governments subsidize burning it for energy, at which point, it becomes rapidly unsustainable, as occurred in Europe recently. By contrast, "nonrenewable" elements such as silver and gold can satisfy the requirements of billions of users if they are recovered and recycled effectively.

Introducing Nutrient Certificates

The concept of Nutrient Certificates as a counterpart to emissions, energy, and carbon trading was first proposed by Katja Hansen as part of her investigations as senior researcher at The Cradle to Cradle Chair, Erasmus University and is further defined here.

> *Definition*. Nutrient Certificates are sets of data describing defined characteristics of materials in products that give them value for recovery and reuse. The certificates are a marketplace mechanism to encourage product designs, material recovery systems, and chain of possession partnerships that improve the quality, value, and security of supply for materials so they can be reused in continuous loops or closed loops or beneficially returned to biological systems. This is done by adding a new value dimension to materials quality. This new dimension is based on the suitability of materials for recovery and reuse as resources in other products and processes.

Nutrient Certificates have a focus distinct from, for example, Environmental Product Declarations (EPD) [35], whose main aim is to catalog the environmental impacts of a product. EPDs are often based on Life Cycle Assessment (LCA) and as a result, face the boundary and scoping uncertainties as well as methodological variations inherent to LCA methods since their inception. Factors such as embodied-energy, transport distances, and "consumption" of resources play a primary role in EPDs. By comparison, Nutrient Certificates focus on describing what is in the product, especially its suitability for reuse in continuous loops or cascade chains and materials pooling.

Nutrient Certificates are also distinct from, but related to, recyclability indexes [36]. However, those indexes focus more on volumes and weights instead of detailed contents. For example, factors such as coatings and additives are often not considered in recycling indexes, and interpretations of "thermal recycling" differ from the Cradle to Cradle® interpretation.

Why the Term "Nutrient Certificate"?

The term "Nutrient Certificate" is used here to signify the fundamental importance of materials as nutrients for other processes. Instead of ending as unusable undefined waste and poor or unusable nutrients as many materials end today, materials would be defined for continuous reuse as nutrients for other processes.

In this case, the term "nutrient" extends beyond the conventional definition applied to a biological nutrient and includes for example rare metals that are "nutrients" for electronics products.

The term "resource certificate" or "materials certificate" could just as easily be applied, but those terms do not recognize the role of resources and materials as nutrients for continuous processes.

Who Could Use Nutrient Certificates?

Various users will benefit from Nutrient Certificates, for example:

- Chemical manufacturers who want to gain access to new markets based on the suitability of their products for use in recyclable materials.
- Materials manufacturers.
- Complex products manufacturers.
- Designers who want to add value to the products they are designing.
- Retailers who want to include recyclability and materials recovery in their purchasing criteria for products they buy.
- Governments who want to provide the marketplace with certainty about how to define materials recycled content and recyclability.
- Recyclers who have to know what is in a product and how it comes apart.
- Builders, building owners, and managers aspiring to go beyond the confines of traditional sustainability.

Intended Applications of Nutrient Certificates

Enhance Quality, Value, and Security in the Chain of Possession

- Value Chain Enhancement.

  Nutrient Certificates would enhance and reinforce the value chain among producers, users, and reprocessors of materials. Those materials used in buildings already have a commercial value when they are sold to contractors and installed in buildings. They also have a residual value as "waste" or used products when they are removed from buildings. Nutrient Certificates provide a mechanism to convert that "waste" into a more defined marketable resource whose residual value is enhanced compared to what it might be today. By so doing, Nutrient Certificates also eliminate the cost of "waste" disposal because there is no waste anymore and instead only resources. For further descriptions of value refer to subsection *Quantifying Economic Value*.

- Authentication.
  Nutrient Certificates when combined with "chain of possession" authentication can protect businesses from industrial counterfeiting, adulteration, and diversion, which cost industry large financial losses annually. In this way, the certificates pay for themselves quickly.
- Security of Supply.
  The continuing supply volatility supply crisis for rare-earth metals suggests that security of supply will be increasingly important to companies. Nutrient Certificates provide a basis for reliable material recovery and pooling in defined pathways that guarantee supplies for manufacturers.
- Transition Mechanism.
  The transformation from low-quality waste to high-quality resources will not occur immediately due to the time frame required for industry to adapt to this new paradigm. Nutrient Certificates can support this transition by identifying materials in a product that can be easily extracted as resources, compared to other portions that might have to be disposed of conventionally. This approach can be described as a "roadmap to improvement" where the percentage of reusable material in a product increases over time. This roadmap can be used along the whole pathway of a material by chemicals manufacturers to improve the recyclability of a virgin material and by manufacturers of complex products to make sure the recycled materials contained in those products are themselves recyclable.
- Value Partnerships.
  A value chain for Nutrient Certificates could be established in a partnership between "waste management" companies, who are already transforming their role into that of "materials managers," and product manufacturers who have a direct relationship with materials suppliers. It is in the financial interests of materials managers to work with product manufacturers to upgrade the recoverability of materials because this improves the value of the materials they reprocess and trade. As well, materials managers have expertise that product manufacturers do not; they know what is required to take apart a product so its contents can be recovered, and they work daily in the materials marketplace with logistics and brokering. It is in the interests of product manufacturers to collaborate with materials managers to redesign products because in this way manufacturers know the chain of possession and in many cases offer greater opportunities to retrieve their materials at a higher level of quality and lower cost through greater ease of disassembly and recovery. Some examples of these are already occurring in the marketplace, for example, with glass, metals, and plastics.

Improve Quality Assurance and Risk Management Standards A quality assurance standard lets companies participate in materials pooling while assuring customers of materials quality. Companies increasingly rely on pools of recycled materials for raw materials, but undefined recycled materials carry inherent risks that limit companies' participation in pooling. In addition, toxic product scandals arising from undefined "virgin" materials have demonstrated that quality assurance is required to manage liability risk. Nutrient Certificates can enhance the quality assurance system by defining conditions for intelligent materials pooling or smart pools. This carefully defined type of materials pooling is known as "intelligent" or "smart," because it is based on information networks that let companies participate in the same way they are beginning to participate in smart energy grids, where multiple input and output points are managed with sophisticated information technologies. Nutrient Certificates could also play a supporting role in other quality assurance systems that require transparency. Cradle to Cradle Certification®, Green building standards, and REACH are examples.

Improve Resource Tracking Between Regions Nutrient Certificates could allow for resources to be tracked as they move between regions, for example, the tracking of phosphate in animal feed flowing from South America to Europe, or rare metals in computer parts flowing from China to assemblers in Japan to customers in Africa. Tracking could serve as a basis for nutrient banking and trading to measure nutrient surpluses and deficits, incentives to restore topsoil used for agricultural production, and mechanisms to compare carbon storage and reuse with carbon depletion.

Nutrient Certificates could provide a basis for a globalized system for more accurately determining available stocks especially of rare materials. A globalized database of Nutrient Certificates would be one way to achieve this. However, such a database is not a prerequisite for starting with Nutrient Certificates. In worst case scenarios, inability to reach accurate estimates of the global circulation of some materials would not discredit the Nutrient Certificate scheme, because unlike emissions mechanisms these are not issued by governments on the basis of estimates, but instead, are issued based on defined materials. In that context, the main shorter-term benefits of Nutrient Certificates are to improve value and reliability in the chain of possession.

Examples of the Contents of Certificates

Many of the parameters prescribed here for Nutrient Certificates are not new, although some, such as *Preferred Ingredients,* were developed as part of Cradle to Cradle® application tools. The critical value of Nutrient Certificates is to provide a way of quantifying and valuing the nutrients each material or product contains. In the following sections, an emerging list of material attributes important for Nutrient Certificates are discussed to stimulate further development.

Material Composition

- *Recycled and/or recyclable content* and the chain by which the material can support recycling or reuse at a high level of quality. This includes a description of *recoverable content,* i.e., what portions are recoverable in a defined recovery pathway and accompanied by disassembly or recovery instructions. In the case of complex products, the disassembly instructions are the more important feature, whereas in the case of component materials the recovery instructions are most important. For example, depolymerizing and repolymerizing of plastics.
- *Defined ingredients* in the material. Aspects such as:
  - Physical properties
  - Element composition
  - Beneficial and harmful off-gassing, leaching, or wearing
  - Elution
  - Stable components
  - Contaminants
  - Thermal degradation products
  - Aging of products where they are transformed over their use period
- *Preferred Ingredients.* Has the material been assessed to determine if it contains preferred safe ingredients? This is a fundamentally different approach than only declaring hazardous ingredients or being “free of” hazardous ingredients.
- Defined Nutrient Classes
  - *Material Classes.* Products are comprised of various classes of biosphere and technosphere nutrients; from basic elements such as copper and silver, to basic materials such as glass and paper, and to the hundreds of thousands of additives that give those materials their functional qualities. Those classes can be defined as follows, but it is emphasized that these provisional definitions are primarily for the purpose of establishing the basis for further definitions rather than core to the Nutrient Certificate quantification or valuation methods. For example, defining a substance as a “chemical” or “compound” will not alter its intrinsic value for Nutrient Certificates.
  - *Base Elements.* Base elements in the periodic table, e.g., copper.
  - *Chemicals.* Basic chemicals used in the manufacture of more sophisticated compounds and materials.
  - *Complex compounds.* Complex combinations, e.g., glucose, of basic chemicals.
  - *Basic Materials.* A basic material can contain multiple elements and compounds, but in this definition, it is normally limited to commodity materials that are broadly used in manufacturing, e.g., plastic, wood, biomass, glass, and cotton, which form components of more complex products.

- *Additives.* Additives are referred to here as chemicals that are added to basic materials to give them functional qualities, e.g., fire retardants in plastics. This can also be applied in the case of, for example, coatings.
- *Products, Complex Products.* This is often a confusing category because a product for one industry is often a chemical or material for another. For example, the chemical industry produces base chemicals as products that are then used in other more sophisticated materials and products. However, for these purposes here, a "product" is defined as something that contains multiple complex chemicals or materials, e.g., specialty concrete, carpet, and furniture. A "complex product" would be a product containing hundreds or thousands of materials, e.g., a computer or vehicle.
- *Grade of Material.* Many national and international grading systems already exist to establish material grades at national and international levels for substances ranging from steel and plastics to topsoil. Those systems can be incorporated into certificates if they provide added valuable information. Nutrient Certificates do not have to reinvent quality standards if they are already available.
- *Biobased, Biodegradable, or Both.* A description of whether the whole material including additives is biobased, and if the material and additives together can be safely biodegraded according to their defined pathways or are to be kept in technical cycles.

Material Functions

Actively Beneficial Functions If a material actively performs an enhancing function such as metabolizing pollutants as some synthetic coatings and natural plants do, or producing electric current via conversion of photovoltaic or kinetic energy. It is important to include such features in a Nutrient Certificate to determine if recycling can maintain such properties.

Material-Defined Pathways

- *Material Pools.* Description of the type of material pools where the material can be used after it is recovered at end of use, and the defined material pools the materials is sourced from, if any.
- *Preferred Defined Use.* "Defined use" establishes if the use of the material in a technological or biological cycle has been defined, and if a preferred pathway in that cycle has been identified. This includes the *Cascade Function.* See the later description of cascades.
- *Chain of Possession.* If the materials come from identifiable sources such as a pool of recycled materials, and how far back along the chain of possession they can be traced.
- *Roadmap to Improvement.* Description of roadmap to improving the product or material over time, if any.
- *Authentication.* If the declaration is authenticated by an outside auditor.

Proprietary Information Although proprietary formulas are sometimes seen as a barrier to transparency, mechanisms still exist for validation, for example, a statement that an independent agency has verified the formulas, and a source where the ingredient information can be obtained. Also, in the case of broadly used materials such as grades of plastics, glass, and steel, there are few if any proprietary barriers for larger volumes.

Industry Labeling Compared to Retail Labeling Nutrient Certificates described here are designed for use by industry and government rather than as consumer product labeling. In the future, certain aspects might enhance product labeling for retail customers, but this is not the intention of Nutrient Certificates.

Certificate Categories There are two distinct categories for Nutrient Certificates, differentiating the biosphere and technosphere metabolisms of materials:

Biosphere Nutrient Certificates Because topsoil and biomass are among the leading terrestrial surface repositories for carbon and have direct interface with atmospheric carbon in a cycle that is essential for agriculture, livestock, and forestry, Biosphere Nutrient Certificates (BNC) could provide a quantifiable and beneficial basis for tracking and trading carbon and other important nutrients. BNC would be applied in cases where materials are intended to be returned to the biosphere during or after use. BNC would apply to products derived from agro-industry and forestry sources, or resulting from post-use biorecovery processes such as biodegradation, biogas generation, topsoil manufacturing, recycled phosphate, and ash recovered from burning of biomass, digestate, or compost.

For example, biosphere Nutrient Certificates could be issued for:

- Food and agro-industry biogenic byproducts
- Digestate from biodigesters
- Manure
- Swill and food from restaurants, households, and other sources
- Fuels
- Paper and paper sludge
- Forestry products
- Biodegradable polymers
- Landscape maintenance clippings
- Ash from biomass incineration
- Sludge from sewage, subject to legislative restrictions
- Phosphate extracted from nutrient streams

Technosphere Nutrient Certificates Technosphere Nutrient Certificates (TNC) provide a quantifiable basis for recovering materials in technical systems, where materials are designed for being used in continuous loops at a similar level of quality.

Technosphere Nutrient Certificates (TNC) could be issued for bulk and commonly used materials such as plastics, steel, aluminum, concrete, and glass, including defined "virgin" as well as "recycled" content, or for more rare materials. For example, TNC could be issued for:

- Base metals, e.g., copper, zinc, silver and rare metals, e.g., gallium
- Defined grades of alloys such as steel
- Defined grades of plastics such as PP, PET
- High quality reusable additives for plastics and metals
- Chemical compounds designed for recovery and reuse

Integrated Nutrient Certificates A complex product containing biosphere and technosphere materials could be accompanied by an integrated certificate describing how biosphere and technosphere components can be separated to become material resources.

Modeling Nutrient Calculations It is not the intention here to describe a detailed model for nutrient calculations. This requires further research into the capacity to quantify each type of nutrient in each type of material. However, as a rule, it is far easier to calculate the actual carbon or nitrogen, phosphorous, potassium, or rare metal content of materials than it is to estimate the "implied" carbon value of many processes, as is the practice with emissions trading today.

Some precedents exist in the C2C methodology for assigning value to nutrients. For example, the textiles company Backhausen in cooperation with Trevira and EPEA has a "Returnity®" certificate [37] that accompanies its products and gives customers a guarantee of 100% recyclability as well as a discount on future purchases if they bring back the product to the manufacturer. The term "Returnity®" embodies the intentions of Nutrient Certificates, returning for eternity.

As well, under C2C certification, a system of "nutrient reutilization scoring" was established as part of a rating system for products [38]. That system, based on an earlier system developed by Braungart et al. and known as the *Intelligent Product System (IPS)* takes into account factors such as Defined Appropriate Cycle, e.g., as technical or biological resources, a documented recovery plan, and "actively closing the loop."

However, the IPS system and C2C certification system are not designed to provide defined information to users and recyclers about features such as material pooling or methods for recovery of nutrients such as phosphate, *in a format that can be attached to materials and tracked along the chain of possession.* Nor are they designed to provide a basis for evaluating the financial recovery value of nutrients in materials.

So, a different type of valuation, authentication, and labeling is required and is further described in the following sections.

Future Directions. Applying and Quantifying Economic Value

Material Cascades In the early sections of this chapter, the unintended economic and nutrient consequences of biomass burning for energy were described. One way to avoid such misguided approaches is to establish "cascades" of material use and reuse.

Cascades can solve various challenges. (1) With many products such as wood or paper, it is not possible to recycle materials at the same level of quality due to deterioration of components during use and recycling. For example, with paper, fibers are damaged and shortened in each reuse until eventually they become unusable. (2) Biomass for paper takes a long time to grow, so if the paper is only used once, then it is incinerated or composted or downgraded quickly, e.g., toilet paper, then the use period quickly exceeds the growth period required for replacement. (3) Immediate incineration of biomass without using it first in products, or after only one use in a product, releases carbon far more quickly than if the materials is reused in products repeatedly.

In material cascades, the goal is to extend the use of the material for as long as practicable instead of using the product once and then burning or composting it [39]. In a cascade, a material can enter technical loops for multiple uses, then return to biological systems as a nutrient or be incinerated with the ash recovered as a nutrient. Alternatively, if a material such as paper or wood cannot be kept in similar quality technical loops, then its downcycling can be extended in a controlled way by recovering high-quality fibers for use in products for as long as possible (see Fig. 3). This type of controlled downcycling is not to be confused with present downcycling where undefined materials are downcycled with minimal knowledge of their content or pathways.

There are quantifiable economic benefits to material cascades. J. Jokinen calculates that when total employment creation is analyzed, the ratio of reusing wood for pulp and paper industry (PPI) products instead of immediately burning it for energy is 13:1 in favor of the PPI alternative [40].

Due to factors like that, the German Federal Ministry of Food, Agriculture, and Consumer Protection has recommended using products/components within the economic system for as long as possible, from a high level of value stepwise to lower levels [41].

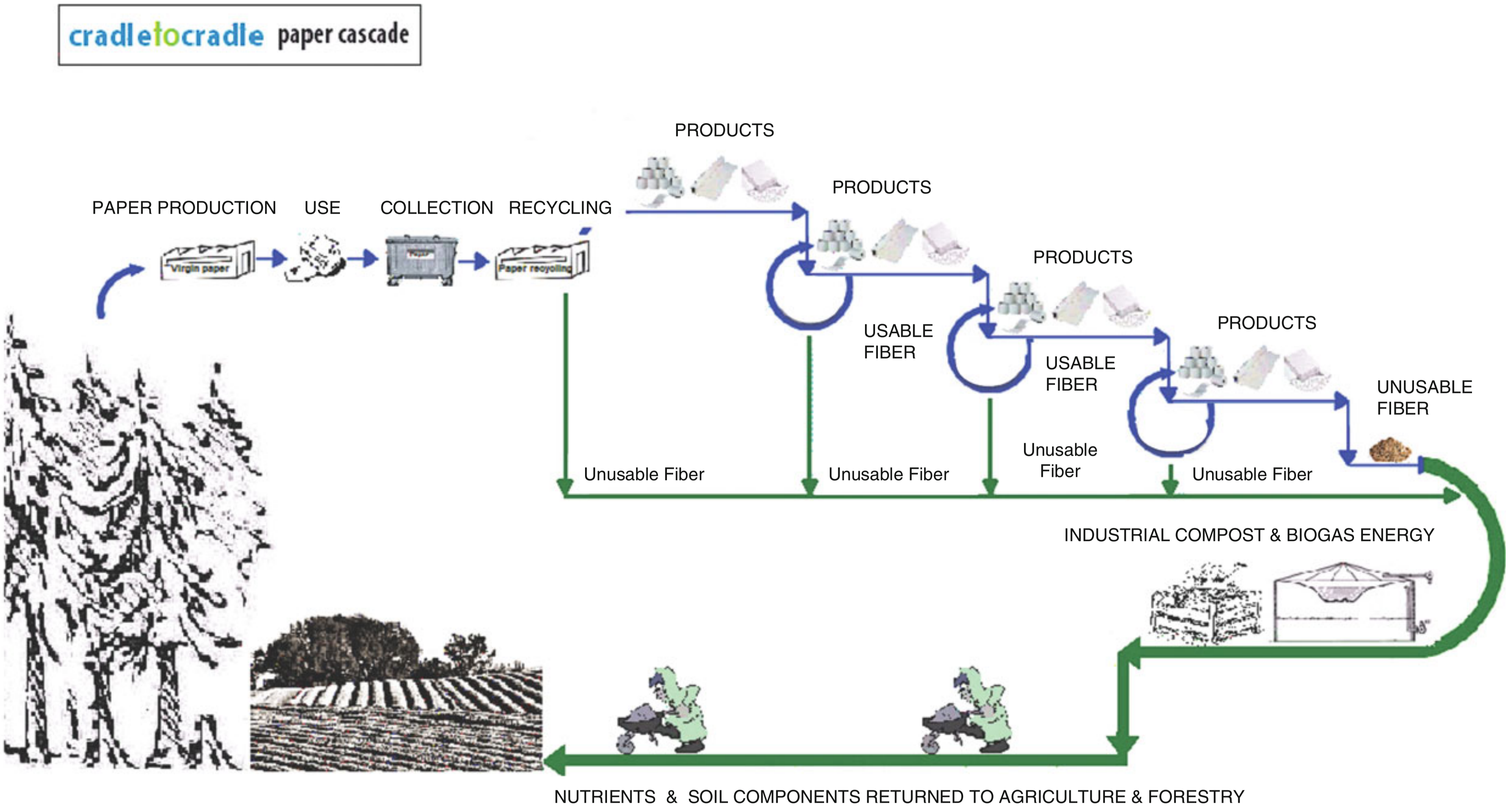
One problem with material cascades is the difficulty of tracing the material as it moves from one use to another. This is especially important to assure that cascades are not misused for downcycling. Nutrient Certificates could be introduced in the cascade process, identifying the cascade supply chain and which stages in the cascade the material has already participated.

Chain of Possession-Tracking Mechanisms It might seem impractical to trace nutrients from bulk materials such as plastics and agro-industry products into more sophisticated complex materials and then into the complex products that may enter a cascade process. However, this form of tracking is already being done to some extent for authenticating mechanisms used to guard against industrial diversion, counterfeiting, and brand abuse [42]. Many companies now offer globalized authentication for a variety of materials ranging from pharmaceuticals to fuel and currencies, with different authentication solutions for different types of materials. Although there is yet no "one size fits all" mechanism to track chain of possession in such a range of materials, the tracking technologies themselves are sufficiently varied and making rapid advances in accuracy as well as affordability to meet this challenge.

Chemical and DNA fingerprinting are advancing quickly as are nano-tagging and electronic micro-tagging, e.g., RFID. These are becoming so sophisticated that they are being used to monitor everything from the origin of oil spills from ships to the origin of uranium in a weapon. Recently, various frameworks have been developed for the management of tagging and tracking systems to support reverse logistics and other chain of possession methods [43]. Large retail chains and manufacturers have sophisticated mechanisms for tracking inventories of component materials used in their products. As well, systems such as Underwriters Laboratory (UL) certifications have been used for years to label and track components in millions of devices. It might be possible to optimize such systems to include Nutrient Certificate-related data.

Building on these systems, a primary implementation path for Nutrient Certificates could be to embed data records in micro-tags that describe "defined-use" standards for recycled and "cascading" raw materials. The data records would complement, not replace, material data safety sheets.

The Potential Role of Buildings for Implementing Nutrient Certificates Nutrient Certificates could be



Resource Repletion, Role of Buildings. Figure 3

Diagram of the potential for using paper in a product cascade (Source: Redrawn from an EPEA drawing [39]) effective mechanisms for transforming buildings and products that move through them to a materials repletion value chain.

Buildings are early candidates for the transition to Nutrient Certificates due to:

- *Value.* Nutrient Certificates represent a way of enhancing the value of buildings for their owners and managers. Instead of materials becoming waste that is expensive to get rid of, they become part of the value chain of the building. This can also enhance property lease and resale value.
- *Tracking.* Buildings involve large material flows that are already quantified in terms of volumes, location, and other specifications. Frequently, inventories are kept of materials used to build, maintain, and operate buildings. These would be transition mechanisms for defining Nutrient Certificate parameters.
- *Chain of Possession.* Institutions such as hospitals, educational institutions, government buildings, large corporations, and airports have extended ownership of their buildings and have a vested interest in maintaining value. They also provide continuity of record-keeping.
- *Volumes.* Building structures use large volumes of materials in one place so it is more convenient to inventory those volumes in one location than trace consumer products that are purchased in small quantities and are transient.
- *Local Material Sourcing.* Many high-volume building materials tend to be locally sourced so it is relatively easy to identify the chain of possession and to go back to the supplier to verify what is in the materials.

Time frame for Recovery of Value from Materials in Buildings Although buildings are perceived as largely longer-term repositories for materials, in reality, many materials used in the construction and operation of buildings have a shorter use period. For example, packaging for construction materials, topsoil removed to make way for new buildings and replaced in the form of landscaping, cleaning products, maintenance equipment, floor coverings, and office equipment all have shorter life cycles. Because of this, Nutrient Certificates can offer economic benefits for building owners and operators early into the process, when materials are transformed from their use in a product to resources for reuse. The resource recovery can occur relatively quickly or may require decades as further discussed below.

Early Candidate Products for Nutrient Certificates in Buildings The “low hanging fruit” for Nutrient Certificates consists of business-to-business materials in buildings that involve a chain of possession where the players are relatively well defined. For example, on the technosphere side, heating, ventilation, and air conditioning (HVAC) equipment and other systems that represent close to half the value of many buildings have defined chains of possession for their component materials, from manufacture through to transport, sales, use, and disposal. Their components are well cataloged for maintenance purposes. “Moveable” office equipments such as photocopiers also have defined chains of possession because they are leased as part of service agreements. On the biosphere side, office paper in large institutions offers chain of possession opportunities because it is collected and reprocessed in large volumes. Interior and exterior landscaping is often maintained as part of service agreements, and these provide a mechanism for quantifying the nutrient contribution of plants and soil. The important characteristic of many of those inventory systems is that they already exist, with accounting systems that can be modified to track and quantify materials. In some cases, individual companies already have a form of Nutrient Certificates for materials used in building interiors, such as the previously discussed Backhausen Returnity® certificates. Less well-defined, but still traceable in large institutions, food waste and sewage can be tracked for organic nutrients.

Early Adopters for Nutrient Certificates in Buildings Because regulatory regimes typically lag behind the science, governments might not be leaders at proactively defining materials. However, governments will be important for advancing recovery of materials as nutrients because they are:

- Good at inventorying what is in their buildings
- Long-term players in the property marketplace
- The owners and occupants of substantial numbers of buildings

Hospitals, educational institutions, large corporations, and airports also are potential early adopters because they have extended ownership of their buildings and a vested interest in maintaining value. One ready-made network of companies exists that is accustomed to the principles and practices underlying a Nutrient Certification system. That is the network of the hundreds of companies working on Cradle to Cradle®-defined products and infrastructures. Nutrient Certificates could be introduced into the supply chain of those companies, involving the thousands of suppliers of those companies as well as millions of customers at the receiving end. In particular, "waste management" companies who are transforming themselves to materials providers are well equipped to implement Nutrient Certificates. For example, in The Netherlands, companies such as Van Gansewinkel Groep have declared their intentions to use C2C application tools in partnerships with customers [45], an important step in bridging the economic intentions of industry with the environmental intentions of customers.

Establishing Bankable and Trading Value Nutrient Certificates do not have to be legislated into existence. Their parameters already provide the marketplace with the motivations, guarantees of content, recyclability, residual value, risk management, and authentication. Governments can accelerate adoption of Nutrient Certificates by specifying them for purchasing, but this is not a prerequisite for their existence.

> *Nutrient Certificates* vs. *Emissions Mechanisms*. One advantage of Nutrient Certificates over emissions trading mechanisms is their connection to traceable materials. When connecting to traceable materials, it is more difficult to distort the marketplace by issuing extra certificates, as occurs with emissions trading. In addition, Nutrient Certificates often quantify what is already given value by the marketplace, so there is no need to match supplies of certificates with, for example, emissions because marketplace mechanisms are already in place to establish value.

Calculating Loss of Value Nutrient Certificates can also be used to calculate *loss of value*. If a material is destroyed, for example through incineration, the certificate is invalidated or modified to reflect the residual value of remaining ash. This procedure is an effective way to measure the true cost of incineration because, presently, the loss of value of millions of tons of materials is not calculated when they are incinerated.

Risk Management Features Nutrient Certificates might on the surface seem to add bureaucracy and costs for the building industry, but in reality, they can do the opposite. By improving the quality of materials, industry can use Nutrient Certificates to:

- Gain added protection from the substantial health liabilities that are arising related to indoor air quality and groundwater pollution. Costs of indoor air pollution to human health have the potential to cripple industries financially.
- Add new value to buildings by enhancing the secondary value of the component materials, which today are often regarded as low value or toxic waste. This "value" can produce benefits only a few years into the operation of a building, when maintenance requirements result in the replacement of damaged and worn out products such as carpeting, lighting, and HVAC components.
- Add new revenue streams from short-term flows of materials through a building. For example, food "waste" from cafeterias and restaurants can be part of materials pooling for industrial composting and biodigestion.
- Add new value to buildings through beneficial functions such as topsoil manufacturing, oxygen manufacturing, CO_2 reuse, energy generation, and recovery of scarce materials.
- More rapidly, amortize the cost of generating electricity by replacing traditional cladding materials with energy-generating materials whose qualities for recycling are defined.
- Cut capital and operating costs through service agreements with companies who lease instead of sell materials and products ranging from office equipment to carpets and power generators.
- Gain emissions credit by defining the greenhouse gas "counter-footprint" of a building more precisely and expanding that counter-footprint to include innovative materials such as cladding materials to generate renewable energy.
- Reducing Risks of Market Distortions. Unlike emissions trading, Nutrient Certificates will make it

more difficult to distort markets. When someone acquires a Nutrient Certificate, they will usually be able to check that the material comes along with it, and vice versa.

DBFMO Risk Management New lifecycle costing financial instruments are conducive to the materials-banking approach. For example, in The Netherlands, government buildings are being financed according to a design-build-finance-maintain-operate (DBFMO) approach where a consortium performs each of those functions and is involved in the building for 15–25 years. This provides an extended planning perspective for designing and recovering materials that move through buildings, and an alternative to the typical DBFMO approach of building-in a financial "cushion" to their bids to guard against future unknown cost overruns. Defined materials and related revenue streams can reduce this uncertainty by providing extra value cushions. For example, if the consortium knows it will not have toxic or other waste management costs and has a lower risk of liability from indoor air pollution, this provides a greater level of certainty. If cladding materials are producing revenues from energy generation, the energy benefits will protect against uncontrolled energy price fluctuations.

In those ways, defined materials provide reliability and predictability in the marketplace.

Conclusions

Innovative quality-based approaches including "defined" materials, cascades, and Nutrient Certificates can be used to overcome unintended consequences of traditional approaches to sustainability such as loss of rare materials due to minimization, undefined recycled content, confusion between biobased and biodegradable, depletion from burning virgin biomass, and distortion of markets due to poorly designed emissions credits schemes. The conservatism of the building industry can be advantageous in applying a quality-based approach due to the industry's emphasis on inventorying and value generation. Buildings can thus transcend the current paradigm of materials depletion to become beneficial materials repletion contributors.

Acknowledgments

Portions of this article regarding criteria for the built environment are excerpted from *Cradle to Cradle® Criteria for the Built Environment*, Mulhall & Braungart, CEO Media 2010, Rotterdam, The Netherlands. Reprinted by permission. http://www.duurzaamgebouwd.nl/bookstore. The authors' appreciation goes to Yael Steinberg and other EPEA scientists for their input.

Bibliography

Primary Literature

1. Coto-Millan P, Mateo-Mantecón I, Domenech Quesada JL, Carballo Panela A, Pesquera MA (2010) Evaluation of port externalities: the ecological footprint of port authorities. In: Coto-Millan P et al (eds) Essays on port economics, contributions to economics. Springer, Berlin/Heidelberg, pp 323–340. doi:10.1007/978-3-7908-2425-4_20, Table 1
2. Chini A (2003) Deconstruction and materials reuse. CIB publication 287. In: Proceedings of the 11th Rinker international conference, Gainesville, 7–10 May 2003
3. Bradsher K (2010) China said to widen its embargo of minerals. *New York Times*, 19 Oct 2010
4. Dempsey J (2010) Decline in rare-earth exports rattles Germany. *New York Times*, 19 Oct 2010
5. European Commission Ad-hoc Working Group on Defining Critical Raw Materials (2010) Critical raw materials for the EU. Report of the Ad-hoc working group on defining critical raw materials, European Commission Enterprise & Industry, Brussels, June 2010
6. Greimel H (2009) Enough lithium for hybrid boom? Most say yes. Automotive News, 21 Sept 2009
7. Pannekoek G, de Bruijne G, Smit B (2010) Phosphorus depletion: the invisible crisis. DPRN Phase II report no. 18
8. Richardson M (2010) China's chokehold on rare-earth minerals. *International Herald Tribune*: 9, 11 Oct 2010
9. Doggett T (2010) U.S. aims to end China's rare earth metals monopoly. Reuters, 30 Sept 2010. http://www.reuters.com/article/idUSTRE68T68T20100930. Accessed 20 Oct 2011
10. Keeley G (2008) Barcelona forced to import emergency water. *The Guardian*, 14 May 2008. http://www.guardian.co.uk/world/2008/may/14/spain.water. Accessed 20 Oct 2010
11. Ahrends A, Burgess ND, Milledge SAH, Bulling MT, Fisher B, Smart JCR, Clarke GP, Mhorok BE, Lewis SL (2010) Predictable waves of sequential forest degradation and biodiversity loss spreading from an African city. Proc Natl Acad Sci USA 107:14556–14561
12. Draft topical outline science and technology of the sustainable built environment. Email from Springer Encyclopedia to Prof. Michael Braungart, 30 Nov 2009

13. Cohen D (2007) Earth's natural wealth: an audit. New Scientist 2605:34–41 (23 May 2007)
14. Bradshaw CJA, Giam X, Sodhi NS (2010) Evaluating the relative environmental impact of countries. PLoS One 5(5):e10440. doi:10.1371/journal.pone.0010440
15. USGBC (2008) LEED 2009 for existing buildings and operations maintenance. United States Green Building Council member approved, Nov 2008
16. DGBC (2010) BREEAM-Nl 2010 label for sustainable real estate, assessor manual new buildings. Dutch Green Building Council Version 2.0, Sept 2010
17. An example of this approach can be seen in USGBC (2008) Innovation in design credit catalog. USGBC, Washington, DC, Mar 2008
18. KPMG (2010) Biomass set to overtake wind as renewable energy champion. In: Powering Ahead: 2010 – an outlook for renewable energy M&A, KPMG survey of global renewable energy mergers & acquisitions, KPMG Cooperative, Switzerland
19. Mantau U, Steierer F, Hetsch S, Prins K (2007) Wood resources availability and demands – implications of renewable energy policies. UNECE, FAO, University of Hamburg, 19 Oct 2007
20. Olsson O (2009) European bioenergy markets: integration and price convergence. Licentiate thesis, Swedish University of Agricultural Sciences, Uppsala
21. United Nations, Economic Commission for Europe (2010) Forest products annual market review 2009–2010. Geneva timber and forest study papers no. 25
22. European Commission (2008) Climate change – can soil make a difference? Report on the conference, Brussels, 12 June 2008
23. Pimentel D, Harvey C, Resosudarmo P, Sinclair K, Kurz D, McNair M, Crist S, Shpritz L, Fitton L, Saffouri R, Blair R (1995) Environmental and economic costs of soil erosion and conservation benefits. Science 267:1117–1123, Estimates of topsoil loss vary but the severity has been well accepted for decades.
24. Brenner J, Paustian K, Bluhm G, Cipra J, Easter M, Elliott T, Kautza T, Killian K, Schuler J, Williams S (2001) Quantifying the change in greenhouse gas emissions due to natural resource conservation practice application in Iowa. The Iowa carbon storage project. Final report to the Iowa conservation partnership, Mar 2001. USDA Natural Resources Conservation Service and Colorado State University, Natural Resource Ecology Laboratory, Fort Collins
25. Kratz S, Schnug E (2006) Rock phosphates and P fertilizers as sources of U contamination in agricultural soils. In: Uranium in the environment. Springer, Berlin/Heidelberg, pp 57–67. doi:10.1007/3-540-28367-6_5
26. Gilbertson T, Reyes O (2009) Carbon trading how it works and why it fails, Critical currents, Dag Hammarskjöld foundation occasional paper series no. 7, Nov 2009
27. Hartesveldt RJ, Harvey HT, Shellhammer HS, Stecker RE (1975) The giant sequoia of the sierra nevada. U.S. Department of the Interior National Park Service, Washington, DC
28. Herva M, Hernando R, Carrasco EF, Roca E (2010) The term "offset" has frequently been applied to emissions or habitat and usually refers to the purchase of offset credits for energy or greenhouse gasses. The term "counter-footprint" to describe some positive impacts of activities is used for example in: Methodological advances in ecological footprinting. In: Bastianoni S (ed) The State of the art in ecological footprint theory and applications, short communications, pp 61–62
29. Cradle to Cradle is a registered wordmark of McDonough Braungart Design Chemistry www.mbdc.com. Accessed 21 Nov 2011
30. www.epea.com. Accessed 20 Oct 2010
31. Mulhall D, Braungart M (2010) Cradle to cradle criteria for the built environment, CEO Media, Rotterdam, Oct 2010
32. McDonough W, Braungart M (2002) Cradle to cradle. Remaking the way we make things. North Point Press, New York
33. McDonough W, Braungart M et al (1992) The hannover principles: design for sustainability. W. McDonough Architects, Charlottesville
34. Duivesteijn A (2008) The Almere principles; for an ecologically, socially and economically sustainable future of Almere 2030, Nieuwe 's-Gravelandseweg 3. Thoth Press, Bussum. ISBN 10; 9068684841
35. Magerholm Fet A, Skaar C, Michelsen O (2009) Product category rules and environmental product declarations as tools to promote sustainable products: experiences from a case study of furniture production. Clean Technol Environ Policy 11(2):201–207. doi:10.1007/s10098-008-0163-6
36. Villalba G, Segarra M, Chimenos JM, Espiell F (2004) Using the recyclability index of materials as a tool for design for disassembly. Ecol Econ 50:195–200
37. Global innovation: RETURNITY® The fabric of many lives. http://www.backhausen.com. Accessed 20 Oct 2010. See also Backhausen Returnity Factsheet and Returnity Info Folder downloadable at same website.
38. Cradle to Cradle® Certification Program, Version 2.1.1, MBDC updated Jan 2010
39. CO2-Speicherung und Wertschöpfung – Holznutzung in einer Kaskade (2009) EPEA Internationale Umweltforschung GmbH, Hamburg, May 2009
40. Jokinen J (2006) Value added and employment in PPI and energy alternative. Study prepared for CEPI by Pöyry Forest Industry Consulting Oy & Foreco Oy, November 2006
41. Bekanntmachung ueber die Förderung der angewandten Forschung auf dem Gebiet der nachwachsenden Rohstoffe im Rahmen des Förderprogramms "Nachwachsende Rohstoffe" der Bundesregierung zum Schwerpunkt "Innovative Mehrfachnutzung von nachwachsenden Rohstoffen, Bioraffinerien" (2008) Bundesministerium fuer Ernährung, Landwirtschaft und Verbraucherschutz, 24 Apr 2008
42. A glossary of industry authentication terms can be found for example at http://www.authentix.com/faqs_terms.asp. Accessed 15 Oct 2010
43. Hans C, Hribernik KA, Thoben K-D (2010) Improving reverse logistics processes using item-level product lifecycle management. Int J Prod Lifecycle Manag 4(4):338–359, 22

44. Brand Stewart (1994) How buildings learn: what happens after they're built. Stewart Brand Viking, New York. This diagram courtesy William McDonough & Partners is a rendition of an earlier published description.
45. Cradle to cradle roadmap, Van Gansewinkel Groep annual report for the year 2009

River Fate and Transport

Zhen-Gang Ji
Minerals Management Service, Herndon, VA, USA

Article Outline

Glossary
Definition of the Subject and Its Importance
Introduction
Fate and Decay
Transport in a River
River Modeling
Future Directions
Bibliography

Glossary

Advection The horizontal transport by flows that move patches of material around but do not significantly distort or dilute them.

Biodegradation The breakdown of a compound by enzyme-mediated transformation primarily due to bacteria, and to a lesser extent, fungi.

Dispersion The mixing of water properties in rivers.

Henry's law A law which states that at a given temperature, the solubility of a gas is proportional to the pressure of the gas directly above the water.

Hydrograph A graph showing time variation in flow rate or stage (depth) of water in a river.

Hydrolysis The reaction of a chemical with water in which splitting of a molecular bond occurs in the chemical and there is formation of a new bond with either the hydrogen component (H^+) or the hydroxyl component (OH^-) of a water molecule.

Manning equation An empirical formulation relating velocity (or flow rate) depth, slope, and a channel roughness coefficient in a river.

Mineralization The process by which a dissolved organic substance is converted to dissolved inorganic form.

Nonpoint SOURCE A pollution source that cannot be traced to a specific spot.

Photolysis The transformation of a compound that results directly from the adsorption of light energy.

Point source A pollution source that comes from a specific identifiable source such as a pipe.

Residence time The time required by a particle to cross a river reach.

River A naturally flowing waterbody.

Volatilization The process representing a chemical substance entering the atmosphere by evaporation from water.

Definition of the Subject and Its Importance

Rivers are naturally flowing waterbodies. Small rivers are also called streams or brooks. Rivers are a watershed's self-formed gutter system and usually empty into an ocean, lake, or another river. This chapter describes the characteristics of rivers and the fate and transport in rivers. The mathematical description of river processes and the modeling of rivers are also described here.

Rivers are complex and dynamic. A river often acts as a sink for contaminants discharged along the river, such as effluents from wastewater treatment plants that discharge nutrients, heavy metals, and/or pathogens into the river. Rivers may also act as sources of contaminants in the watershed, depending on the time of the year or the section of the river. The health of a river is directly linked to the health of the surrounding watershed. The water quality in a river will deteriorate, if the watershed condition deteriorates. Via rivers, pollutants can travel hundreds or even thousands of kilometers and cause environmental problems in a waterbody that is located far away from the sources. The common designated uses of a river include aquatic life support, water supply, and recreation activities (such as swimming, fishing, and boating).

Introduction

The rivers and their tributaries, normally occupying less than a few percent of the total drainage basin, are the conduits of the river basin. They are like a gutter system and transport water, nutrients, sediment, and toxicants downstream (often to an estuary or a large lake).

Compared with lakes and estuaries, the most distinct characteristic of a river is its natural downstream flow. Lakes typically have much smaller flow velocities than rivers. Flow velocities in estuaries, though their magnitudes can be comparable to the ones in rivers, are tidally driven and can be in either direction (downstream or upstream).

The origins of contaminants can be divided into point and nonpoint sources. Point source pollution comes from a specific, identifiable source such as a pipe. Nonpoint source pollution cannot be traced to a specific spot. Point sources include wastewater treatment plants, overflows from combined sanitary and storm sewers, and industry discharges. Nonpoint sources include runoffs from urban, agriculture, and mining areas. Point and nonpoint sources have caused a wide range of water quality problems and the deterioration of the ecological state in rivers. Leading pollutants and stressors in the USA include [1]:

1. Pathogens (bacteria)
2. Siltation
3. Habitat alterations
4. Oxygen-depleting substances
5. Nutrients
6. Thermal modifications
7. Toxic metals
8. Flow alterations

Pathogens are the most common pollutant affecting rivers and streams in the USA. Pathogen pollution is a major public health problem especially in the use of river water for water supply and the consumption of fish and shellfish harvested in rivers and estuaries. Bacteria commonly enter surface waters in inadequately treated sewage, fecal material from wildlife, and runoff from pastures, feedlots, and urban areas.

Sediment siltation is one of the leading environmental problems in rivers. The filling of river channels, harbors, and estuaries by sediments brings a high cost to society. The condition of a river's watershed greatly affects the amount of sediment delivered into the river. The sediment sources vary among rivers, and even within a particular river, from year to year. Extreme events, such as hurricanes, can produce dramatic changes in the amounts and types of sediments that are delivered into a river. The vulnerability of a river to sediments and contamination reflects a complex combination of upstream flows, land use, and land-management practices. The vast majority of river sediments is discharged during only 10% of the year (36 days), and 90% of the year represents a very small amount of the sediment load [2]. Low flow rates usually result in net deposition conditions. High flow rates may cause net erosion in upstream reaches and net deposition in downstream reaches or in the estuary into which the river flows.

Often water quality is defined in terms of concentrations of the various dissolved and suspended substances in the water, for example, temperature, salinity, dissolved oxygen, nutrients, phytoplankton, bacteria, heavy metals, etc. The distribution of these substances can be calculated by a mathematical model. Based on the principle of conservation of mass, the concentration change can be represented simply in a one-dimensional form [3]:

$$\frac{\partial \mathrm{C}}{\partial \mathrm{t}} = -\mathrm{U}\frac{\partial \mathrm{C}}{\partial \mathrm{x}} + \frac{\partial}{\partial \mathrm{x}}\left(D\frac{\partial \mathrm{C}}{\partial \mathrm{x}}\right) + \mathrm{S} + \mathrm{R} + \mathrm{Q} \qquad (1)$$

where C = substance concentration
t = time
x = distance
U = advection velocity in x direction
D = mixing and dispersion coefficient
S = sources and sinks due to settling and resuspension
R = reactivity of chemical and biological processes
Q = external loadings to the aquatic system from point and nonpoint sources

The changes of concentration C in Eq. 1 are determined by the following:

1. The hydrodynamic processes control the water depth (D), the advection (represented by the U term), and mixing (represented by the D term).
2. The size and properties of sediment (or particular organic matter) affect the settling and resuspension (represented by the S term).
3. The chemical and biological reactions of pathogens, toxics, and/or nutrients are represented by the R term.
4. External loadings from point and nonpoint sources are included by the Q term.

Fate and Decay

Contaminants in rivers include nutrients, organic toxicants, heavy metals, and pathogens. If no degradation reactions occurred in Nature, every single contaminant discharged in the past would still be polluting the environment. Fortunately, natural purification processes dilute, transport, remove, and degrade contaminants. It is essential to understand the kinetics of reactants and to describe them mathematically. This section summarizes the fate and decay of contaminants and their mathematical formulations.

The fate and transport of contaminants are controlled by two factors: their reactivity and their hydrodynamic transport. Reactivity includes:

1. Chemical processes
2. Biological processes
3. Bio-uptakes

Transport in a river, which will be discussed in the next section, includes three mass transport processes:

1. Advection of water current
2. Diffusion and turbulent mixing within the water column
3. Deposition and resuspension on the water-sediment bed interface

Mathematical Formulations

How long contaminants remain in a waterbody depends on the nature of the compound. Most chemicals undergo chemical or biological decay. Some chemicals are conservative and do not undergo these types of reactions, even though it is very difficult to find a truly conservative chemical in Nature. The fate and decay of a contaminant represent the gradual decrease in the amount of a substance in a river, as the result of various sink processes, including chemical and biological transformation, or dissipation/deposition to other environmental systems.

Although reaction kinetics in aquatic systems can be described in numerous ways, the form for a single reactant is generally expressed as:

$$\frac{dC}{dt} = \mathrm{R} = -\mathrm{k}C^{m} \tag{2}$$

where m = the order of reaction

k = rate constant of the m-order reaction

In natural waters, the commonly used forms of Eq. 2 are with m = 0, 1, and 2.

Zero-order reactions: A zero-order reaction (m = 0) represents irreversible degradation of a reactant that is independent of the reactant concentration. The solution to Eq. 2 is:

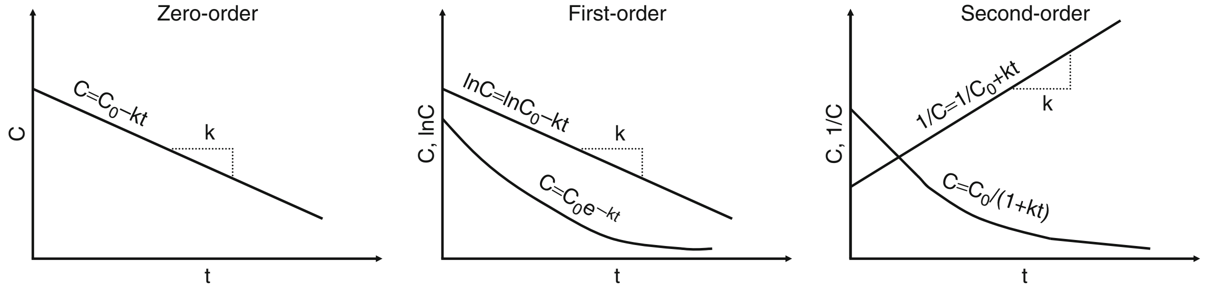
$$\mathrm{C} = \mathrm{C}_0 - \mathrm{kt} \tag{3}$$

where C_0 = the initial concentration at t = 0. In this case, a plot of concentration versus time should yield a straight line with a slope of k, as shown in the left panel of Fig. 1. Zero-order reactions have their reaction rates determined by some factor other than the concentration of the reacting materials.

First-order reactions: First-order reactions (m = 1) have their reaction rates proportional to the concentration of the reactant and are most commonly used in describing chemical and biological reactions. For first-order reactions, the solution to Eq. 2 is:

$$\mathrm{C} = \mathrm{C}_0 e^{-\mathrm{kt}} \tag{4}$$

Equation 4 indicates that for first-order reactions, reactant concentration decreases exponentially with



River Fate and Transport. Figure 1
Left panel: concentration versus time for zero-order reaction. *Middle panel*: concentration and logarithm concentration versus time for first-order reaction. *Right panel*: concentration and inverse concentration versus time for second-order reaction

time. In this case, a plot of logarithm concentration versus time should yield a straight line with a slope of k, as shown in the middle panel of Fig. 1. Most of the reactions found in the environment can be conveniently expressed by a first-order approximation without much error. Examples of first-order reactions include biochemical oxygen demand in surface waters, death and respiration rates for bacteria, and production reaction of algae.

Second-order reactions: For second-order reactions (m = 2), the solution to Eq. 2 is:

$$\frac{1}{C} = \frac{1}{C_0} + kt \tag{5}$$

Therefore, if a reaction is indeed second-order, a plot of inverse concentration of C (1/C) with time should yield a straight line with a slope of k (the right panel of Fig. 1). Equation 5 can also be expressed as

$$C = \frac{C_0}{1 + kC_0 t} \tag{6}$$

which reveals that, similar to the first-order reaction, the resulting concentration of a second-order reaction also decreases and approaches zero as time increases.

Processes Affecting Fate and Decay

The fate and decay of contaminants can result from physical, chemical, and/or biological reactions. In addition to sorption and desorption, processes that can significantly affect the fate and decay processes include:

1. Mineralization and decomposition
2. Hydrolysis
3. Photolysis
4. Biodegradation
5. Bioconcentration
6. Volatilization

Most decay processes are expressed as first-order reactions. The first-order decay coefficients for individual processes are additive and can be linearly superimposed to form a net decay coefficient:

$$k_d = k_m + k_h + k_p + k_{bd} + k_{bc} + k_v \tag{7}$$

where k_d = net decay coefficient
k_m = mineralization coefficient
k_h = hydrolysis coefficient
k_p = photolysis coefficient
k_{bd} = biodegradation coefficient
k_{bc} = bioconcentration coefficient
k_v = volatilization coefficient

In modeling studies, either the net degradation coefficient or the individual coefficients can be specified.

Mineralization and Decomposition Mineralization is the process by which a dissolved organic substance is converted to dissolved inorganic form. Mineralization makes nutrients, such as nitrogen and phosphorus, available for a fresh cycle of plant growth. Bacteria decompose organic material to obtain energy for growth. Plant residue is broken down into glucose that is then converted to energy:

$$C_6H_{12}O_6 + O_2 \xrightarrow{\textit{energy released}} CO_2 + H_2O \tag{8}$$

In water quality models, the term “mineralization” often represents the process by which dissolved organic matter is converted to dissolved inorganic form, and thus includes both heterotrophic respiration of dissolved organic carbon and mineralization of dissolved organic phosphorus and nitrogen.

Hydrolysis Hydrolysis is the reaction of a chemical with water, in which splitting of a molecular bond occurs in the chemical and there is formation of a new bond with either the hydrogen component (H^+) or the hydroxyl component (OH^-) of a water molecule. This involves ionization of the water as well as splitting of the compound hydrolyzed:

$$RX + H_2O \rightarrow ROH + HX \tag{9}$$

Essentially, water enters a polar location on a molecule and inserts itself, with an H^+ component going to one part of the parent molecule and an OH^- component going to the other. The two components then separate. The concentration of hydrogen and hydroxide ions, and therefore pH, is often an important factor in assessing the rate of a hydrolysis reaction. Hydrolysis is a major pathway for the degradation of many toxic organics.

Photolysis Photolysis is the transformation of a compound that results directly from the adsorption of light energy. Compounds that absorb sunlight may gain sufficient energy to initiate a chemical reaction. Some of these photochemical reactions result in the decomposition or transformation of a substance.

The energy of light varies inversely with its wavelength. Longwave light lacks sufficient energy to break chemical bonds. Short wave light (x-rays and gamma rays) is very destructive. Fortunately for life on earth, this type of radiation largely is removed by the upper atmosphere. Light near the visible spectrum reaches the earth's surface and can break the bonds of many organic compounds, which can be important in the decay of organic chemicals in a water system.

The basic characteristics of photolysis are:

1. Photolysis has two types of energy absorption: direct photolysis and indirect photolysis. The direct photolysis is the result of direct absorption of sunlight by the toxic chemical molecule. Indirect photolysis is the result of energy transfer to the toxic chemical from some other molecule that has absorbed the sunlight.
2. Photolysis is the destruction of a compound activated by the light energy and is an irreversible decay process.
3. Products of photolysis may remain toxic and the photolysis process does not necessarily lead to detoxification of the system.
4. The photolysis coefficient in Eq. 7 is usually a function of the quantity and wavelength distribution of incident light, the light adsorption characteristics of the compound, and the efficiency at which absorbed light produces a chemical reaction.

Biodegradation Biodegradation is the breakdown of a compound by enzyme-mediated transformation, primarily due to bacteria, and to a lesser extent, fungi. Although these types of microbial transformations can detoxify and mineralize toxics, they can also activate potential toxics. The rate of biodegradation can be very rapid, which means that biodegradation is often one of the most important transformation processes in rivers.

Even though the biodegradation process is largely mediated by bacteria, the growth kinetics of the bacteria is complicated and is not well understood. As a result, toxic models often assume constant decay rates rather than modeling the bacteria activity directly. The first-order decay rate is commonly used. Biodegradation rate is influenced by water temperature and can be represented by an Arrhenius function:

$$k_b = k_{b20}\theta^{(T-20)} \tag{10}$$

where k_b = biodegradation rate

k_{b20} = biodegradation rate at 20°C

T = water temperature in °C

θ = temperature correction factor

The effect of the Arrhenius function is that a higher temperature will cause a faster chemical reaction rate. It gives a quantitative relationship between the reaction rate and its temperature.

Biodegradation rate is also related to the contaminant concentration and can be expressed by a typical Michaelis–Menten formulation:

$$k_b = k_{bmax}\frac{c}{c + c_{1/2}} \tag{11}$$

where k_{bmax} = the maximum biodegradation rate

c = the contaminant concentration

$c_{1/2}$ = half saturation (Michaelis) constant.

The combination of the above two formulations yields

$$k_b = k_{max}\theta^{(T-20)}\frac{c}{c + c_{1/2}} \tag{12}$$

where k_{max} = maximum decay rate due to biodegradation. Equation 12 combines the effects of contaminant concentration and water temperature on the biodegradation process.

Volatilization Volatilization represents a chemical substance entering the atmosphere by evaporation from water. Volatilization is often treated as an irreversible decay process, because of its mathematical similarities to these decay processes. However, volatilization is actually a reversible transfer, in which the dissolved concentration in water attempts to equilibrate with the gas phase concentration in the overlying atmosphere. Equilibrium occurs when the partial pressure exerted by the chemical in water equals the partial pressure of the chemical in the atmosphere.

Henry's law states that, at a given temperature, the solubility of a gas is proportional to the pressure of the gas directly above the water. Volatilization is often treated similarly to surface oxygen exchange, where the volatilization flux is proportional to the difference between the chemical concentration in water and the saturation concentration, as:

$$F_v = k_v(c_w - c_{ws}) \tag{13}$$

where F_v = volatilization flux

k_v = transfer rate

c_w = dissolved concentration of the chemical in water

c_{ws} = saturation dissolved concentration of the chemical in water

Equation 13 indicates that the chemical enters the water when the chemical in the water is unsaturated ($c_w < c_{ws}$) and the chemical leaves (volatizes from) the water when the chemical in the water is oversaturated ($c_w > c_{ws}$). The saturation dissolved concentration is dependent upon the atmospheric partial pressure and Henry's law constant for the chemical. The transfer rate, k_v, depends on the properties of the chemical as well as the characteristics of the waterbody and the atmosphere, including the molecular diffusion coefficient of the chemical in the water and in the atmosphere, the temperature, the wind speed, the current velocity, and the water depth.

Transport in a River

Rivers have distinct hydrodynamic characteristics that are different from those of lakes or estuaries. This section focuses on the following:

1. River flow and the Manning equation
2. Advection and dispersion processes in rivers

River Flow and the Manning Equation

The flow rate of a river is the volume of water that passes a cross section of the river in a unit of time, which is usually expressed in cubic meters per second (cms) or cubic feet per second (cfs) and is calculated as:

$$Q = A\ V \tag{14}$$

where Q = Flow rate in cms or cfs

A = Area through which the water is flowing in m^2 or ft^2

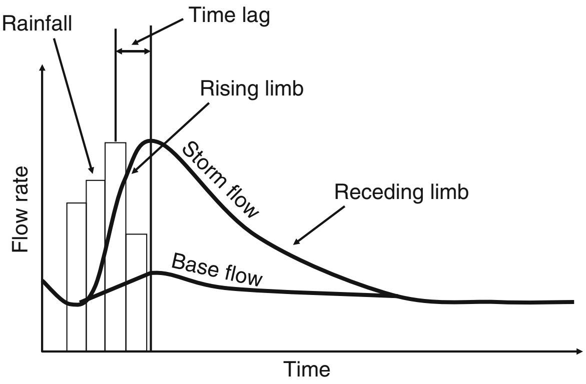
V = Average velocity in the downstream direction in m/s or ft/s

The river flow can generally be separated into two components:

1. Base flow
2. Storm flow

Base flow is composed largely of groundwater effluent and sustains river flow during dry weather periods. Storm flow is from the runoff during or shortly after a precipitation event. The water from base flow is the precipitation that percolates into the ground and flows slowly through a long path before reaching the river, whereas the water from storm flow is the precipitation that reaches the river shortly after precipitation through runoff. In addition to base flow from the groundwater and the storm flow from the runoff, point sources, such as wastewater treatment plant discharges and tributaries to the river, also contribute to a river flow.

A hydrograph is a graph showing time variation in flow rate or stage (depth) of water in a river. As sketched in Fig. 2, the river flow is composed of the storm flow and the base flow. After the beginning of a rainfall, the storm flow from runoff starts to increase and reaches its peak some time after the peak rainfall. There is a time lag between the two peaks. The rising limb is the portion of the hydrograph to the left of the peak of the storm flow, which shows how long the river takes to reach its peak flow rate after a rainfall event. The receding limb is the portion of the hydrograph to the right of the peak, which shows how long the river takes to return to the base flow.



River Fate and Transport. Figure 2
A storm hydrograph of a river

In addition to flood events, low flow conditions are also important characteristics of a river. When there is no precipitation contributing to the storm flow, and the base flow from groundwater is low, the river experiences low flow conditions. Low flow results in less water available for dilution of pollutants from point sources, causing high pollutant concentrations in the river. Therefore, point source discharges during low flow conditions have the most significant impact on the river's water quality, since the discharge may constitute a larger percentage of river flow. For instance, wastewater discharges to the Blackstone River can account for up to 80% of the total river flow in summer [4].

A hydrodynamic model based on momentum and continuity equations is often used to calculate flow velocity, flow rate, and water depth in a waterbody. A simpler approach to calculate these parameters is to use the Manning equation, which is an empirical formulation relating velocity (or flow rate), depth, slope, and a channel roughness coefficient in a river. The Manning equation was derived by curve-fitting data measured in rivers and channels. The equation is:

$$V = \frac{Q}{A} = \frac{R^{\frac{2}{3}} S^{\frac{1}{2}}}{n} \tag{15}$$

where V = mean flow velocity in m/s

Q = flow rate in m^3/s

A = cross-sectional area in m^2

R = hydraulic radius in m

S = slope of the channel bed in m/m

n = Manning roughness coefficient

The hydraulic radius is defined as:

$$R = \frac{A}{P} \tag{16}$$

where P is the wetted perimeter in m, which is the length of contact of the water with the channel in m, measured in a direction normal to the flow. The Manning roughness coefficient, n, represents the channel roughness that contributes to the dissipation of flow energy. Table 1 shows a range of n values for various channels and rivers.

Originally developed in the 1880s, the Manning equation is still widely used in hydraulic calculations with reasonable accuracy today. In hydrodynamic modeling, the Manning equation may serve the purpose of giving a quick estimation of flow conditions in a river. However, the Manning equation is an empirical formulation that may not reflect actual conditions of a river.

River Fate and Transport. Table 1 Values of the Manning roughness coefficient, n, for various channels and rivers [5]

Type of channel	Manning roughness coefficient (n)
Smooth concrete	0.012
Ordinary concrete lining	0.013
Earth channels in best condition	0.017
Straight unlined earth canals in good condition	0.020
Natural rivers and canals	0.020–0.035
Mountain streams with rocky beds and rivers with variable sections and some vegetation along banks	0.040–0.050
Alluvial channels without vegetation	0.011–0.035

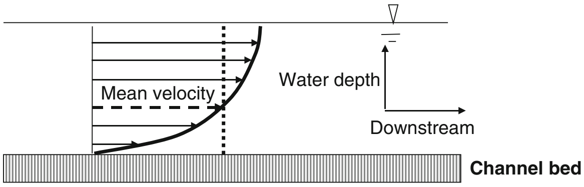
Advection and Dispersion in Rivers

Advection refers to horizontal transport by flows that move patches of material around but do not significantly distort or dilute them. In rivers, advection often represents the primary transport process of pollutant in the longitudinal direction. Dispersion is the mixing of water properties. In rivers, a prominent feature is the longitudinal dispersion: the transport and spreading of pollutants downstream from a point source. When a tracer is released into a river, two distinct processes control the tracer transport:

1. Flow advection carries the tracer away from the releasing point.
2. Turbulence dispersion spreads out and dilutes the tracer concentration.

Mathematically, the above two processes are represented by the first and second terms on the right-hand side of Eq. 1, respectively. Advection results in the pollutant's moving downstream, while longitudinal mixing leads to spreading or smearing in the longitudinal dimension. Lateral and vertical mixing processes determine how long it takes for a pollutant to be completely mixed across a river. The dominant transport process in rivers is the advection due to river flow. Flow velocity controls a river's residence time, the time required by a particle to cross a river reach. The dispersion process in rivers is often less important in the transport of pollutants. The effect of dispersion may be ignored in analyzing a continuous pollutant load to a river. Figure 3 is a velocity vertical profile in a channel. In small rivers, however, the turbulence generated by bed friction is strong, and the depth is generally small, resulting in rivers that are often well mixed vertically.

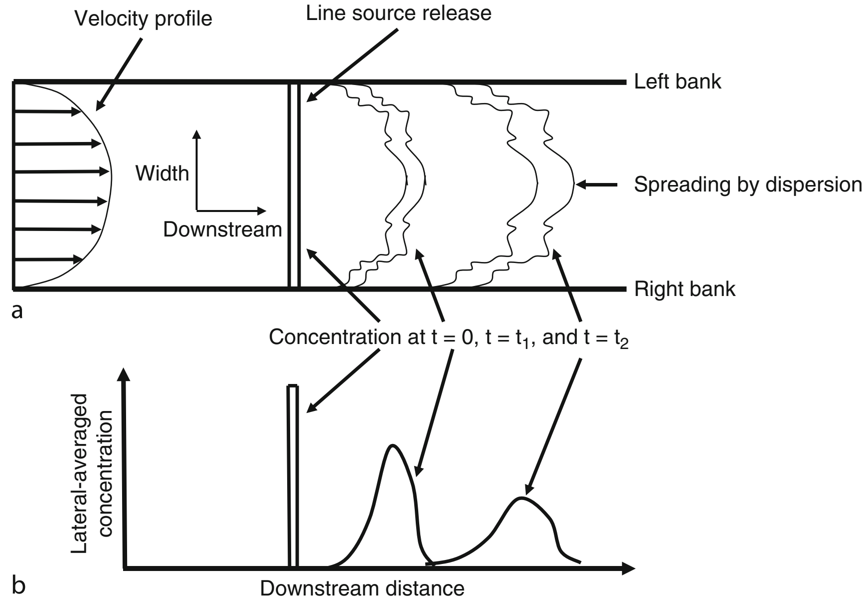
To illustrate the longitudinal dispersion in a river, an idealized dye release experiment is shown in Fig. 4, in which Panel A gives the plain view of the dye transport in the river and Panel B presents the lateral-averaged dye concentration along the river. In the river, a line source of constant concentration is instantaneously released at time t = 0, and the longitudinal velocity has parabolic variation across the river. As shown in Panel A, the advection process transports the dye downstream, and the dispersion process spreads the dye and reduces the maximum concentration. Dye travels downstream faster in the middle of the river than near the banks. As a result, the line source released at t = 0 becomes approximately a parabolic shape at t = t_1 and t = t_2. The concentration profiles at t_1 and t_2 in Panel A also reflect the random fluctuations of turbulence activities in the river. Because of variations in flow velocity across the river, dye spreads both along and across the river by dispersion. The laterally averaged dye concentration in Panel B also indicates that the velocity shear and turbulent dispersion contribute to the concentration spreading along the river.



River Fate and Transport. Figure 3
Velocity vertical profile in a channel

Impacts of River Flow on Water Quality

Water quality processes can be highly dependent on river flow conditions. The time that a pollutant remains



River Fate and Transport. Figure 4
Advection and dispersion processes in a river. *Panel A* gives the plain view of dye transport in the river. *Panel B* presents the lateral-averaged dye concentration along the river

within a section of a river is called residence time. The flow velocity and the length of the river section determine the residence time. River flow affects water quality in a river in several ways:

1. *Dilution.* A large volume of flow dilutes concentrations of pollutants that are discharged into the river.
2. *Residence time.* High flow velocity reduces the residence time and affects the amount of material that can be produced or degraded in the river section.
3. *Mixing.* High flow velocity increases mixing in the river, enhances the assimilative capability of the river, and reduces pollutant concentration gradients.
4. *Erosion.* High flow can erode bed material and destabilize the benthic environment.

The impact of pollutant loadings to a river is largely determined by the magnitudes of the loadings and the flow rate. Rapid transport of pollutants by high flow results in a short residence time and often causes minimal water quality problems. Conversely, slow transport of pollutants by low flow results in a long residence time and can lead to water quality problems, such as oxygen depletion and eutrophication. Channel alteration and watershed disturbance can lead to abnormally high flow rates for a given amount of rain and amplify the impact of floods. Watershed disturbance can also increase sedimentation and harm aquatic biota in a river.

In temperate regions, seasonally high flow typically occurs during the periods of snowmelt in early spring and spring rains, whereas seasonally low flow normally occurs in summer and early fall. The river flow affects the concentration and distribution of water quality variables. Generally, point sources have a larger impact on a river during low flow (dry weather) conditions due to less water diluting the pollutants. Low DO concentrations and high algal growth in a river often occur during low flow periods and hot weather conditions. The combination of low flow, minimum dilution, and high temperature often makes summer and early fall the critical periods for evaluating the impact of point sources (such as wastewater treatment plants).

In contrast, nonpoint sources can bring large amounts of pollutants from the watershed into a river during high flow (wet weather) conditions. It is important to examine both point and nonpoint sources in both high and low flow conditions. Point sources of nutrients often cause algal blooms in rivers during low flow conditions, while nonpoint sources may increase nutrient concentrations and turbidity following periods of wet weather events. Municipal discharges, agriculture runoff, and urban runoff are among the most common sources of impairment to rivers.

In the study of the Blackstone River, for example, Ji et al. [2] reported that a discharge from a wastewater treatment plant was the dominant point source of contaminants and had significant impact on the sediment contamination in the river. However, this point source alone is still insufficient to account for the total metal concentrations in the river. Nonpoint sources and the processes of sediment deposition and resuspension are also important factors that control the concentrations of sediment and toxic metals.

Dissolved oxygen is essential to river ecosystems. Processes controlling DO spatial distribution in a river include:

1. Oxidation of the biochemical oxygen demand (BOD): BOD is used to represent all sinks of dissolved oxygen, such as the oxidation of carbonaceous and nitrogenous organic matter, the benthic oxygen demand, and the oxygen utilized by algal respiration.
2. Reaeration of DO from the atmosphere: In addition to atmospheric reaeration, DO produced by photosynthesis and DO contained in incoming flows are also major oxygen sources.
3. Transport due to the river flow: Advection and diffusion processes enhance DO mixing and reaeration within a river.

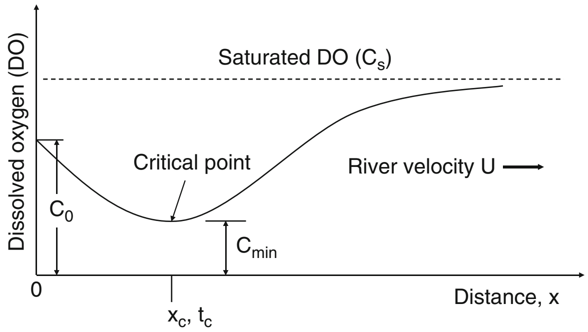
The pioneering work by Streeter and Phelps [6], who developed the first water quality model to describe the oxygen depletion in the Ohio River, is useful for understanding DO processes in a river. It can be described in a first-order reaction equation:

$$U\frac{dC}{dx} = -k_d B + k_a(C_s - C) \tag{17}$$

where x = distance

U = advection velocity in x direction

C = DO concentration



River Fate and Transport. Figure 5
DO sag curve in a river

B = BOD concentration

C_s = saturated dissolved oxygen concentration

k_d = deoxygenation rate constant of BOD

k_a = first-order reaeration rate constant of DO

By assuming that BOD has a first-order degradation reaction with a decay rate constant of k_r, the solution to Eq. 17 is the famous Streeter-Phelps equation:

$$C = C_s - \frac{k_d L_0}{k_a - k_r}\left(e^{-k_r x/U} - e^{-k_a x/U}\right) - (C_s - C_0)e^{-k_a x/U} \qquad (18)$$

A schematic representation of the Streeter-Phelps equation is shown in Fig. 5, describing a DO sag curve in a river. The DO sag curve gives DO longitudinal variation as the result of oxygen depletion and recovery, after a BOD load is discharged into a receiving river. Between the discharge point ($x = 0$) and the critical distance ($x = x_c$), oxidation exceeds reaeration (i.e., $k_d B > k_a(C_s - C)$ in Eq. 17) because of high BOD concentrations and a small DO deficit ($= C_s - C$). Oxygen in the river is consumed faster than it is resupplied. The DO concentration decreases to a minimum C_{min} at a critical distance x_c (or critical time $t_c = x_c/U$). This position is the critical location where the lowest DO concentration occurs, and the oxidation rate and reaeration rate are equal. After passing the critical location, reaeration exceeds oxidation (i.e., $k_d B < k_a(C_s - C)$ in Eq. 17) because of a low BOD concentration and a high DO deficit. Thus, oxygen in a river increases gradually. Further downstream, the rate of supply exceeds the utilization rate, resulting in a full recovery of the DO concentration.

River Modeling

The two primary reasons to conduct river modeling are:

1. To better understand physical, chemical, and biological processes
2. To develop models capable of realistically representing rivers, so that the models can be used to support water quality management and decision making

Water quality management needs to understand key processes affecting environmental problems in order to evaluate management alternatives. Examples of such environmental problems are:

1. Thermal pollution due to power plant discharges
2. Sedimentation in harbors causing siltation and high dredging costs
3. Eutrophication due to excessive nutrient loadings
4. Low dissolved oxygen conditions caused by wastewater discharges
5. Accumulation of toxic materials in the sediment bed

Models play a critical role in advancing the state-of-the-art of hydrodynamics, sediment transport, and water quality, and of water resources management. Because of their requirements for precise and accurate data, models also ultimately contribute to the design of field data collection and serve to identify data gaps in characterizing waterbodies. Models are used to analyze the impact of different management alternatives and to select the ones that result in the least adverse impact to the environment.

Models are often used to improve the scientific basis for theory development, to make and test predictions, and to clarify cause-and-effect relationships between pollutant loadings and the receiving waterbody. Models are often used to evaluate and test potentially expensive water quality management alternatives prior to their implementation. The cost of a hydrodynamic and water quality modeling study is usually a small fraction of the implementation cost. Models can simulate changes in an ecosystem due to changes in internal and/or external conditions, such as water elevation variations or increased external pollutants. These simulations predict positive or negative changes within the river ecosystem due to the management actions, such as improved sewage treatment or reduced agricultural runoff. These simulations are obviously far more cost-effective than testing expensive management actions on a trial-and-error basis, thus making models a useful tool for water quality management.

In the past decades, hydrodynamic and water quality models have evolved from simplified one-dimensional, steady-state models, such as the legendary QUAL2E model [7], to complex three-dimensional, time-dependant models of hydrodynamics, sediment, toxics, and eutrophication. Three-dimensional modeling has matured from a research subject to a practical engineering tool. Over this same period, computational requirements for realistic three-dimensional modeling have changed from supercomputers, to high-end workstations, and then to personal computers.

These advanced three-dimensional and time-dependant models, which can also be readily applied for one- and two-dimensional problem settings, provide a powerful computational tool for sediment transport, water quality, eutrophication, and toxic chemical fate and transport modeling studies. These advanced models often include several coupled submodels for different physical, chemical, biological processes in surface waters, such as:

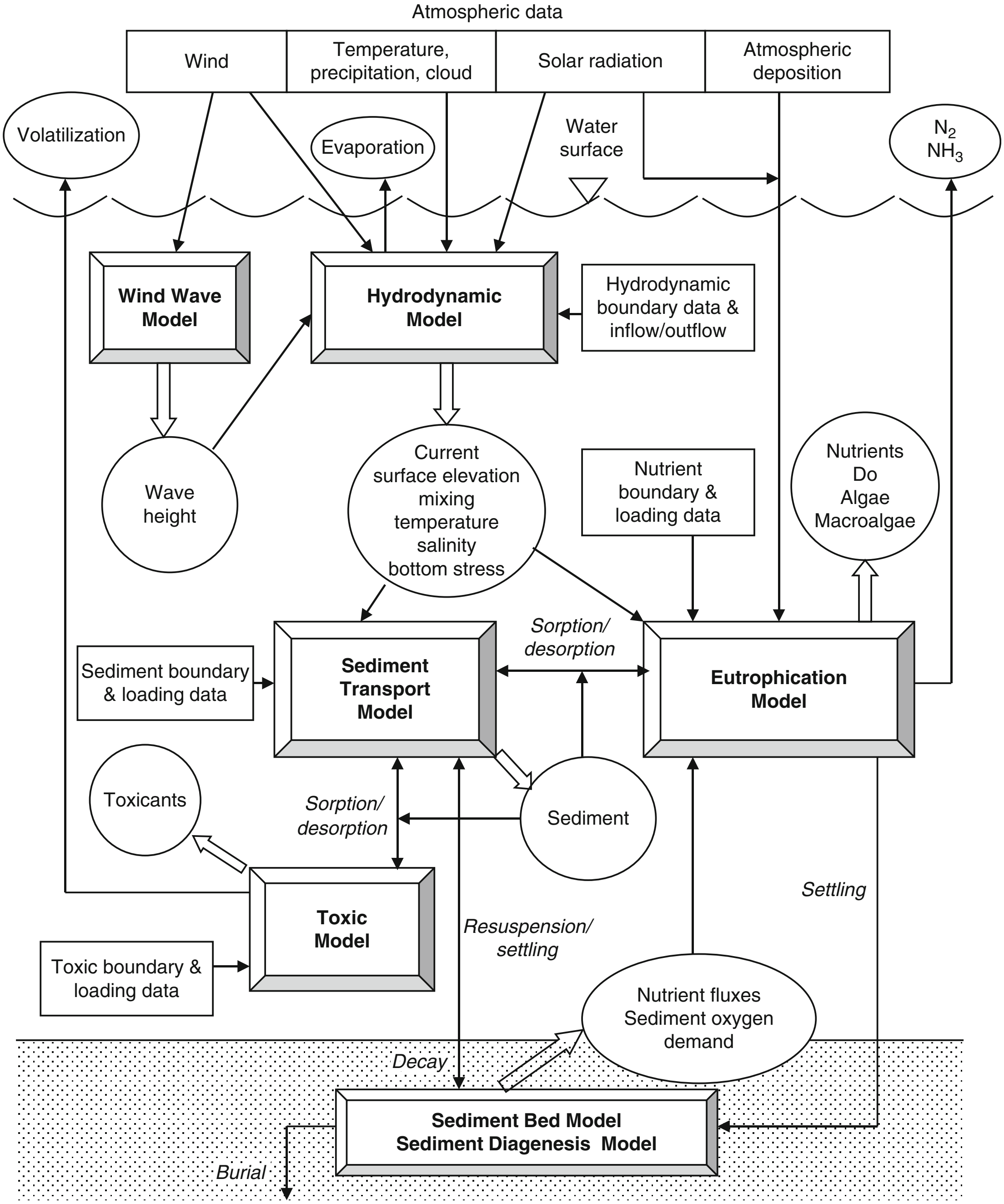
1. Hydrodynamic model
2. Wind wave model
3. Sediment model
4. Toxic model
5. Eutrophication model
6. Sediment diagenesis model
7. Submerged aquatic vegetation (SAV) model

As an example, Fig. 6 illustrates the major components of the Environmental Fluid Dynamics Code (EFDC) model [8]. In addition to computational modules, these advanced models tend to evolve into complex software systems, comprising many tools and sources of information. They may contain components for grid generation, data analysis, preprocessing, postprocessing, statistical analysis, graphics, and other utilities.

Transport in rivers is often dominated by the processes of advection and dispersion. One-, two-, and three-dimensional models have been developed to describe these processes. Study objectives, river characteristics, and data availability are key factors determining model applicability. In river studies, 1D and steady-state models are commonly used, such as the QUAL2E model [7]. If a river is wide enough to have significant lateral variations or deep enough to develop vertical stratifications, 2D (and even 3D) models may be needed to simulate sediment and toxicant transport in the river. For instance, sediment transport within a meandering river is very complex. The velocities are faster at the outer bank and slower at the inner bank. The lateral velocity difference directly influences the sediment transport. There might be erosion occurring along the outer bank and deposition occurring on the inner bank. Using a 1D model to represent the river is equivalent to treating sediment transport as being uniform across the river, eliminating the effect of river meandering on sediment transport and vertical stratifications. A 1D model represents the entire cross section of the river as being either net depositional or net erosional.

One-dimensional models, such as the widely used QUAL2E model [7], are traditionally applied to river modeling. For most small and shallow rivers, these 1D models are often adequate to simulate hydrodynamic and water quality processes. In 1D models, water surface elevation, velocity, and discharge vary only in the longitudinal (along-the-river) direction and are constants in the lateral (across-the-river) direction. This approach provides a simplified mathematical description of river flows.

Rivers with a steep bottom slope often have a relatively high velocity and a shallow water depth, and are characterized by gravel, cobbles, and rocks in the riverbed. Coarse sands and finer particles are washed out by the high velocity. The dominant gradient of water quality constituents is along the river in the direction of flow. A 1D laterally and verticallyaveraged model is thus appropriate for describing water flow and the transport of sediment and toxic chemicals. Rivers with a moderate bottom slope result in a low-velocity waterway, often characterized by a sediment bed consisting of a mixture of fine-grained cohesive particles and fine sands. The dominant gradient of water quality constituents in this kind of river is in the direction of the flow and a 1D model may still be adequate. One-dimensional models are limited in their ability to capture the complexity of natural rivers. The assumption that the characteristics of the river are uniform both vertically and laterally may


Atmospheric data
Wind
Temperature, precipitation, cloud
Solar radiation
Atmospheric deposition
Volatilization
Evaporation
Water surface
N_2 NH_3
Wind Wave Model
Hydrodynamic Model
Hydrodynamic boundary data & inflow/outflow
Wave height
Current surface elevation mixing temperature salinity bottom stress
Nutrient boundary & loading data
Nutrients Do Algae Macroalgae
Sorption/ desorption
Sediment boundary & loading data
Sediment Transport Model
Eutrophication Model
Toxicants
Sorption/ desorption
Sediment
Settling
Toxic Model
Resuspension/ settling
Toxic boundary & loading data
Nutrient fluxes Sediment oxygen demand
Decay
Sediment Bed Model Sediment Diagenesis Model
Burial

River Fate and Transport. Figure 6
Major components (submodels) of the EFDC model

not be valid for wide, deep rivers. In this case, the 1D approach may fall short of describing the river processes. Transport in these rivers can have significant gradients either laterally or vertically. In this case, a 2D or 3D model is needed to provide a better representation of the river. Ji et al. [4] gave an example of modeling hydrodynamics, sediment transport, and toxics in a small, shallow river.

Future Directions

The fate and transport of contaminants in rivers are complicated processes that include physical transport and chemical and biological kinetics. Contaminants in a river may be the result of either past or present disposal practices. Shutting off the sources does not always solve the problem (e.g., DDT persists many years). Consequently, it is essential that mathematical models for assessing contaminants are accurate and reliable. In the past decades, significant progress has been made in numerical model development, data collection, and computer software and hardware. These developments have helped mathematical models to become reliable tools for environmental management and engineering applications.

"Modeling is a little like art in the words of Pablo Picasso. It is never completely realistic; it is never the truth. But it contains enough of the truth, hopefully, and enough realism to gain understanding about environment systems" [9]. Water quality management increasingly depends upon accurate modeling. This dependency is further amplified by the adoption of the watershed-based approach to pollution control. Models enable decision-makers to select better, more scientifically defensible choices among alternatives for river water quality management.

In many cases, the models are used to evaluate which alternative will be most effective in solving a long-term water quality problem. The management decisions require the consideration of existing conditions, as well as the projection of anticipated future changes of the water system. In these applications, the river models not only need to represent the existing conditions, but also have to be predictive and give conditions which do not yet exist. Models are also used to provide a basis for economic analysis, so that decision makers can use the model results to evaluate the environmental significance of a project as well as the cost-benefit ratio.

Three key factors have contributed to the great progress in the modeling of rivers:

1. Better understanding and mathematical descriptions of physical, chemical, and biological processes in rivers
2. Availability of fast and efficient numerical schemes
3. Progress in computer technology

The powerful, yet affordable computers in combination with fast numerical algorithms have enabled the development of sophisticated 3D hydrodynamic and water quality models. These advanced models contain very few simplifying approximations to the governing equations. Personal computers (PCs) have evolved rapidly to become the standard platform for most engineering applications (with the exception of very large scale problems). PCs represent the most widely used computer platform today. Models developed on a PC can be transformed to other PCs without much difficulty. The relatively low prices of PCs also make modeling more cost-effective. Due to the rapid advances in computer technology, PCs are now widely used in river modeling studies.

Bibliography

Primary Literature

1. USEPA (2000) National water quality inventory: 1998 report to Congress. EPA 841-R-00-001. US Environmental Protection Agency, Office of Water, Washington, DC
2. CSCRMDE (1987) Sedimentation control to reduce maintenance dredging of navigational facilities in estuaries. Report and symposium proceedings. Committee on sedimentation control to reduce maintenance dredging in estuaries, National Academy Press, Washington, DC
3. Ji Z-G (2008) Hydrodynamics and water quality: modeling rivers, lakes, and estuaries. Wiley, Hoboken, 676 pp
4. Ji Z-G, Hamrick JH, Pagenkopf J (2002) Sediment and metals modeling in shallow river. J Environ Eng 128:105–119
5. Chow V (1964) Handbook of applied hydrology, a comparison of water-resources technology. McGraw Hill, New York
6. Streeter HW, Phelps EB (1925) A study of the pollution and natural purification of the Ohio river. III: factors concerned in the phenomena of oxidation and reaeration. Bulletin Number 146, US Public Health Service
7. Brown LC, Barnwell TO (1987) The enhanced stream water quality models QUAL2E and QUAL2E-UNCAS: documentation and user manual. EPA/600/3-87-007. US Environmental Protection Agency, Athens, Georgia
8. Hamrick JM (1992) A three-dimensional environmental fluid dynamics computer code: theoretical and computational aspects. The College of William and Mary, Virginia Institute of Marine Science, Special Report 317, p 63
9. Schnoor JL (1996) Environmental modeling: fate and transport of pollutants in water, air, and soil. Wiley, New York

Books and Reviews

Blumberg AF, Mellor GL (1987) A description of a three-dimensional coastal ocean circulation model. In: Heaps NS (ed)

Three-dimensional coastal ocean models, coastal and estuarine science, vol 4. American Geophysical Union, Washington, DC, pp 1–19

Bowie GL, Mills WB, Porcella DB, Campbell CL, Pagenkopf JR, Rupp GL, Johnson KM, Chan PWH, Gherini SA (1985) Rates, constants, and kinetics formulations in surface water quality modeling, 2nd edn. USEPA, Environmental Research Laboratory, Athens. EPA/600/3-85/040

Casulli V, Cheng RT (1992) Semi-implicit finite difference methods for three-dimensional shallow water flow. Int J Numer Meth Fl 15:629–648

Cerco CF (1999) Eutrophication models of the future. J Environ Eng 125(3):209–210

Chapra SC (1997) Surface water-quality modeling. McGraw-Hill, New York, 844 pp

Chapra SC, Canale RP (1998) Numerical methods for engineers, with programming and scientific applications. McGraw-Hill, New York, 839 pp

Di Toro DM (2001) Sediment flux modeling. Wiley, New York

Fischer HB, List EJ, Imberger J, Brooks NH (1979) Mixing in inland and coastal waters. Academic, New York, 483 pp

Gill AE (1982) Atmosphere-ocean dynamics. Academic, New York, 662 pp

Hutchinson GE (1957) A treatise on Limnology. In: Geography, physics and chemistry, vol I. Wiley, New York, p 1015

Ji Z-G (2004) Use of physical sciences in support of environmental management. Environ Manage 34(2):159–169

Ji Z-G (2005) Water quality models: chemical principles. In: Water encyclopedia, vol 2, Water quality and resources development. Wiley, New Jersey, pp 269–273

Ji Z-G (2005) Water quality modeling-case studies. In: Water encyclopedia, vol 2, Water quality and resources development. Wiley, New Jersey, pp 255–263

Martin JL, McCutcheon SC (1999) Hydrodynamics and transport for water quality modeling. Lewis, Boca Raton

Morel F (1983) Principles of aquatic chemistry. Wiley, New York, 446 pp

Park K, Kuo AY, Shen J, Hamrick JM (1995) A three-dimensional hydrodynamic-eutrophication model (HEM3D): description of water quality and sediment processes submodels. The College of William and Mary, Virginia Institute of Marine Science. Special Report 327, 113 pp

Schumm SA (1977) The fluvial system. Wiley, New York

Thomann RV, Mueller JA (1987) Principles of surface water quality modeling and control. Harper and Row, New York

USEPA (1994) Water quality standards handbook, 2nd edn. US Environmental Protection Agency, Office of Water, Washington, DC, EPA-823-B-94-005b

USEPA (1998) Bacteria water quality standard status report. US Environmental Protection Agency, Office of Water, Washington, DC

USEPA (2000) Nutrient criteria technical guidance manual: rivers and streams. EPA-822-B-00-002. Office of Water, Office of Science and Technology, Washington, DC

Wezernak CT, Gannon JJ (1968) Evaluation of nitrification in streams. J Sanit Eng Div ASCE 94(SA5):883–895

Wool AT, Ambrose RB, Martin JL, Corner EA (2003) Water quality analysis simulation program (WASP), Version 6: Draft users manual. Available at: http://www.epa.gov/athens/wwqtsc/html/wasp.html

Ziegler CK, Nesbitt B (1994) Fine-grained sediment transport in Pawtuxet river, Rhode Island. J Hydraul Eng 120:561–576

Ziegler CK, Nesbitt B (1995) Long-term simulation of fine-grained sediment transport in large reservoir. J Hydraul Eng 121:773–781

Roots and Uptake of Water and Nutrients

Carvalho P., M. J. Foulkes
Division of Plant and Crop Sciences, School of Biosciences, University of Nottingham, Loughborough, UK

Article Outline

Glossary
Definition of the Subject and Its Importance
Introduction
Root System Morphology and Anatomy
Effects of Abiotic Stress and Root Development
Root Traits and Resource Capture
Application of Rooting Traits in Breeding for Tolerance of Abiotic Stresses
Future Directions
Bibliography

Glossary

Fibrous root system Root system formed by various root axis of similar size, typical of cereals.

Lateral roots Lateral roots are the roots formed from the pericycle cells of other roots. The first-order laterals refer to the roots emerging from the primary and secondary root axes. Second-order laterals emerge from the first-order laterals, and third-order laterals from the second-order lateral, and so on. Usually, lateral branching is limited to the fifth-order laterals.

Primary roots Often called seminal roots, these are the first root axes to develop arising from the coleorhizae of the seed.

Rhizosphere Volume of soil immediately adjacent to plant roots (usually between 10 and 20 mm), which is affected by their growth, secretions, respiration, nutrient and water, and associated soil microorganisms.

Root architecture Describes the spatial configuration of the root system as a whole. Since it describes multiple root axes it subsumes both topology and distribution.

Root cap Root cap is the tissue that covers the apex of the root. It protects the apical meristem, acts as gravisensor tissue, and facilitates the passage of the growing roots by producing root mucilage.

Root distribution Root distribution refers to the distribution of different root traits, often morphologic ones (e.g., weight, length, volume), as a function of several factors, the most common being soil depth.

Root hair Specialized projection formed by a modified epidermal root cell. It augments the total surface area of a root system, dramatically increasing its absorption capacity.

Root morphology Root morphology refers to the surface features of a single root axes as an organ. It includes the characteristics of the epidermis such as root hairs, root cap, pattern of appearance of lateral roots, cortical senescence, and diameter. Weight, volume, and area are also part of the morphology.

Secondary roots Secondary roots are the roots that grow from the hypocotyl, the coleoptile, stem, and tillers; they are also called crown, nodal, or adventitious roots. This term is also used to describe the roots emerging from the primary roots; however, first-order laterals is a better term to describe those roots.

Taproot In many gymnosperms and dicotyledons, the primary root axis to arise from the seed, greatly enlarges to become the most prominent root axis of the plant, and is usually referred to as a taproot.

Taproot system Taproot system refers to the root systems formed from a central and usually relatively large root axis, the taproot.

Topology Describes the branching pattern of the individual root axes.

The main quantitative root traits and resource capture variables discussed in this entry are summarized in Table 1.

Definition of the Subject and Its Importance

The UN forecasts that the world population will reach 9.4 billion by 2050. The world must therefore develop the capacity to feed 10 billion within the next 40–50 years [1, 2]. The increase in production has to come from greater yields on existing cropland; but also without proportionate increases in the use of water or fertilizer, and within the context of climate change [3, 4]. A substantial increase in the effectiveness with which available water and nutrients are used is therefore required to ensure food security and environmental protection in future decades. Water is recognized as the most limiting factor in crop production worldwide, though nutrient shortages may often be as important as water scarcity and worldwide recovery of nitrogen (N) fertilizer in cereal systems worldwide is on average low at ca. 30–50% [5], with implications for potential environmental impacts. Crop improvement under these conditions seems likely to be increasingly dependent on breeding for deeper or denser root systems, which promote soil moisture and nutrient capture and high dry matter production in cultivars subjected to water and/or nutrient stresses. In this entry, information is set out on the current understanding of the structure and functions of crop root systems. The avenues for the optimization of root anatomy and morphology traits that could be applied to the genetic and agronomic improvement of crop root systems for more effective below-ground resource capture are considered. Specifically, the determinants of effectiveness of capture by root systems of two key resources, water and nitrogen, according to their structure and function are considered.

Introduction

All higher plants have roots and the root fraction of the plant's total dry weight varies widely, both between and within species [6]. Although roots encounter many environmental fluctuations that affect their growth, they have a capacity to adjust to these as a whole system that makes them strongly dynamic in their spatial and temporal expansion. Understanding of these modifications in the form and function of the root system and their relationships with resource capture, whether due to environmental response or genetic control, is of importance for sustainable crop production.

Roots and Uptake of Water and Nutrients. Table 1 List of physiological variables, their definitions and units

Physiological variable (symbol)	Definition	Units
Nitrogen-use efficiency (NUE)	Grain DM at harvest (kg) per N available (soil + fertilizer) (kg)	Dimensionless
Nitrogen-uptake efficiency (UPE)	Above-ground N (kg) at harvest per N available (soil + fertilizer) (kg).	Dimensionless
Water-use efficiency (WUE)	Above-ground DM per water used (soil evaporation + crop transpiration).	$g\ l^{-1}$
Transpiration efficiency (TE)	Above-ground DM (g) per water used (l) through crop transpiration.	$g\ l^{-1}$
Radiation-use efficiency (RUE)	Above-ground dry mass (g) per intercepted global radiation (MJ).	$g\ MJ^{-1}$
Root to shoot ratio (R:S)	Ratio between the root and above-ground shoot dry mass.	Dimensionless ($g\ g^{-1}$)
Root mass ratio (RMR)	Ratio between the root dry mass (g) and total plant dry mass (g).	Dimensionless ($g\ g^{-1}$)
Root mass (RM)	Dry mass (g) of the total root system.	g
Root length (RL)	Total length (cm) of all roots present.	cm
Root length density (RLD)	Root length (cm) per unit of soil volume (cm^3).	$cm\ cm^{-3}$
Root volume (RV)	Total volume of the root system (cm^3).	cm^3
Root diameter	Average diameter (mm) of an individual root; commonly assumed to be a cylinder.	mm
Root fineness (RL:RV)	Ratio of the total root length (mm) to total root volume (mm^3).	$mm\ mm^{-3}$
Root depth	Maximum depth reached by a plant root system (m)	m
β Parameter	Describes the cumulative root distribution with depth. Estimated from: $p = 1 - \beta^d$, where p is the proportion of roots accumulated from the surface to a depth, d.	Dimensionless
k Parameter	Resource capture coefficient. Estimated from $\Phi = 1 - e^{-k.RLD}$, where Φ is the proportional resource captured (resource captured/ available resource).	m^2
Specific root length (SRL)	Ratio between root length (m) and root weight (g).	$m\ g^{-1}$
Root tissue density (RM:RV)	Ratio between root dry mass (mg) and root volume (mm^3).	$mg\ mm^{-3}$
Root front velocity (RFV)	Downward extension rate of the root front (mm) per day.	$mm\ day^{-1}$
Root longevity	Length of time of which an individual root is present in the soil.	days

Although the influence of canopy characteristics on above-ground productivity of crops is now relatively well understood, due to the difficulty of access and complexity of environment interactions, understanding the role of the root system is less complete. There is no doubt of the importance of the form and function of root systems to water and nutrient capture, and information has increased in the last decades on the role of crop root systems in maintaining yields under abiotic stress [7–9]. Root traits are a relatively new target in crop improvement programs aimed at improving tolerance of abiotic stress (water and nutrients). There are reports that particular root morphology and/or anatomical traits help plants maintain higher grain yields under low resource availability, for example, relatively deeper distribution of roots increasing water uptake under drought in wheat [10] and rice [11], longer root hairs increasing P acquisition under low P availability in barley [12], and narrower root xylem vessels in wheat were associated with increased water uptake during grain filling [13]. Nevertheless, the genetic control of root characteristics is poorly understood in most major staple crops, especially in bread wheat. Future improvement will depend on a better understanding of the morphological and anatomical traits determining below-ground resource capture, as well as the development and application of phenotypic screens to characterize genetic variation in the key traits. In this entry, the prospects for manipulating roots systems for improved resource capture and yield (drought and low nutrient availability) are considered, with a particular emphasis on cereal crops. The root traits that are focused on are principally morphological relating to root proliferation, root biomass and root length density, and their distribution with depth; although some consideration is given to anatomical features, for example, xylem root frequency and diameter. Roots and the uptake of water and nutrients are considered with regard to: (1) root morphology, (2) responses of root systems to water and nutrient stress, (3) the capacity of root systems for resource capture, and (4) prospects for breeding crops with optimized root systems for resilience to abiotic stresses.

Root System Morphology and Anatomy

Due to the difficulty of access and complexity of environmental interactions, roots are still one of the most challenging subjects in plant investigations, but their importance is unquestionable. Anchorage, support, and water and nutrient uptake are the main functions of the plant root system. With regard to crop root systems, the terms morphology and architecture are frequently used. The root system may be characterized according to four main categories of morphology/architecture [14, 15]. "Root morphology" refers to the surface features of a single root axes as an organ. It includes the characteristics of the epidermis such as root hairs, root cap, pattern of appearance of lateral roots, cortical senescence, and diameter. Root weight, volume, and area are also part of the morphology. "Root topology" describes the branching pattern of the individual root axes. "Root distribution" refers to the distribution of root traits, often morphologic ones (e.g., biomass length, biomass, etc.) as a function of several factors, the most common being soil depth. Finally, "root architecture" relates to the spatial configuration of the root system as a whole.

The root morphology of monocotyledons differs from that of dicotyledons in several important respects. In the monocotyledons, for example, small grain cereals, two types of roots constitute the root system: the primary and the secondary roots [7, 16]. The primary roots (often called seminal roots; usually between three and eight axes) develop first arising from the coleorhizae of the seed [17, 18] and are active throughout all the crop life cycle [18]. Their extension is mainly downward allowing them to occupy the deeper layers of the soil profile [17]. The secondary (often called crown, nodal, or adventitious roots) are the roots that grow from the nodes of the coleoptile, main shoot, and tillers. The onset of tillering is the starting point of the growth of the secondary roots, and their formation is intimately related to tiller formation [19], so that factors favoring tillering will increase secondary root production. In dicotyledons, for example, oilseed rape, for most species the primary root consists of a taproot from which lateral roots and their branches arise. In both monocotyledons and dicotyledons, lateral roots are initiated in the pericycle and grow through the cortex to emerge at the surface of the parent root.

Rooting depth is affected by root-penetration rate and phenology. Generally, the longer a crop is growing, the deeper it roots [20]. Rooting depth (maximum depth reached by the roots) determines the amount of the soil that a plant can explore. Maximum rooting depth is typically 140–200 cm in winter cereals [18, 21] and 80–120 cm in spring cereals [22]. Rooting depth also strongly depends on the soil type and depth as well as below-ground resource availability, but generally the longer a crop grows the deeper the root system. Root growth and rooting depth of crop plants can be restricted because of physical and chemical impediments. Where physical and chemical soil constraints are absent, the maximum depth of rooting on deep soils is genetically determined and differs not only between vegetation types but also between crop species grown under identical conditions [8]. Much of within season variation in maximum rooting depth can be explained by temperature [23].

In monocotyledons, immediately after sowing root growth is favored, followed by a gradual decrease in assimilate partitioning to the root in favor of shoot growth after emergence. After flowering, the aboveground growth (fruit and grain formation) is favored, whereas root weight usually remains constant or decreases [24, 25]. Thus, root biomass and total length production generally follow a sigmoidal pattern from sowing to flowering in cereals, at which point further increases are not usually observed [18, 21].The root dry mass ratio (root DM/total DM; RMR) is ca. 0.3 in wheat and barley during early growth, decreasing to ca. 0.1 at harvest [18, 22]. Another important trait influencing the crop's capacity to capture resources per unit soil volume is the root length density (RLD), which describes the total root length per unit of soil volume. Typical values of RLD in the upper 0.1 m of soil are about 20 cm cm^{-3} in grasses, 5–10 cm cm^{-3} in temperate cereal crops and 1–2 cm cm^{-3} in other crops [8]. The distribution of the RLD through the soil profile typically follows an exponential decrease with depth [26]. The cumulative distribution of RLD with depth (β) can be approximated by the equation described by Gale and Grigal [27] as:

$$Y = 1 - \beta^d \tag{1}$$

where Y is the fraction of the root system accumulated from the soil surface to depth, d, and β is a parameter that describes the shape of the cumulative distribution with depth.

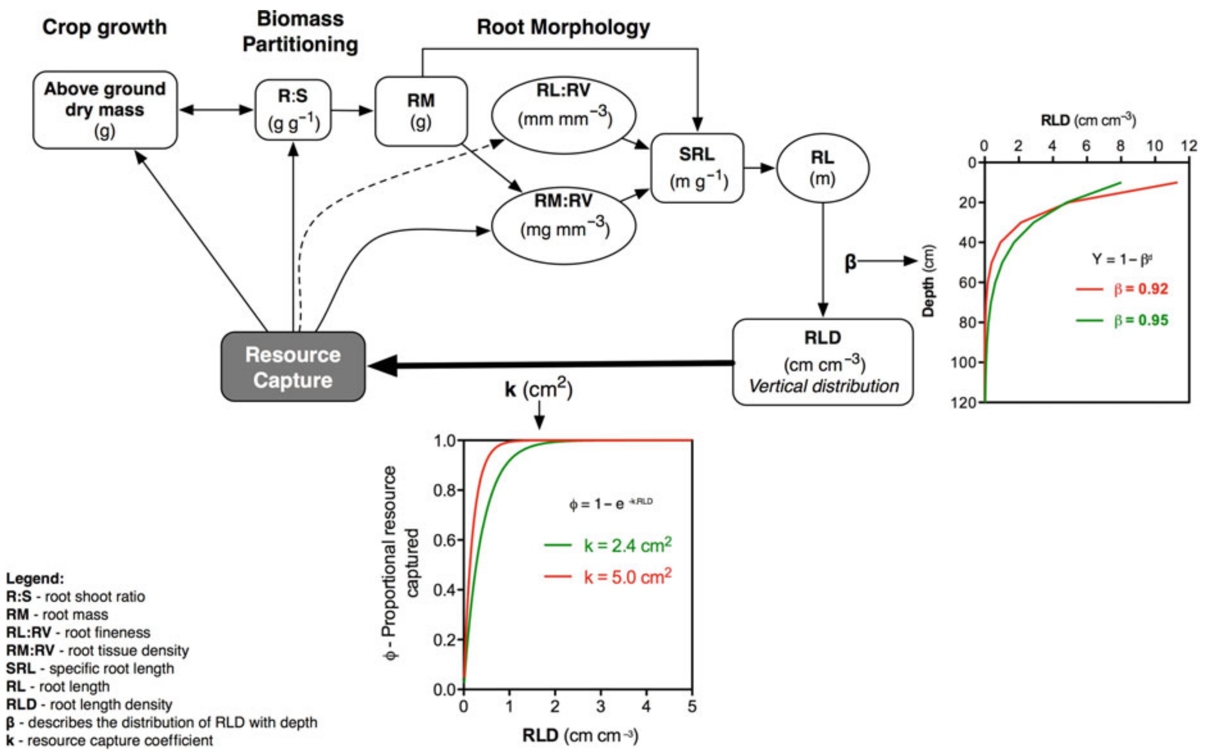
This equation has been widely adopted since (e.g., [28–30]). The distribution of roots of many crops (e.g., cauliflower and winter wheat) is well described by the relationship, but in others (e.g., oilseed rape and sugar beet), this relation is found in the surface layers, but there is a tendency for values of RLD in deeper soil layers to be almost constant [8]. Differences in the distribution of RLD with depth may be associated with the velocity at which roots elongate to depth (root front velocity [RFV]) and the proliferation rate at each soil layer [31]. Root front velocity is closely related with the water and N extracted by the crop [31–33]. Another important trait influencing the potential for water and nutrient acquisition by roots is the mean root diameter (see Root traits and resource capture below).

Specific root length (root length per unit root DM - SRL; km g^{-1}) strongly influences the RLD. Theoretically, a high SRL (thinner roots) would be beneficial especially in resource-deficit situations. Specific root length is also positively correlated with root extension rate [34]. Specific root length varies considerably in the field and is strongly affected by environment; typical values are 130–250 m g^{-1} for cereals [35–37]. Root tissue density (root weight (RW): root volume (RV)) is highly correlated with root life span but inversely correlated with root expansion [38, 39]. So low RW: RV will may be one strategy to increase SRL of crop root systems [40] and potentially resource acquisition. Specific root length is a complex parameter that is determined by root length, tissue density, and diameter. It influences plant investment in potential resource acquisition (RLD) but also reflects root longevity and root growth rate, and therefore it is of potential interest as a selection criteria in breeding programs for optimized root systems. Maximizing SRL seems to be an advantage particularly in water- and nutrient-limited conditions [39, 41] and is associated with higher RLD. Intuitively, thinner roots would be advantageous for acquiring soil resources, though there may be trade-offs with other root functions such as anchorage, support, and transport [42]. The interrelationships between these root traits and their relations with resource capture are summarized in Fig. 1. Additionally, there are reported effects of root diameter on resource capture, which has been shown to be highly correlated with plant dry mass [43, 44] and the diameter of conducting vessels. The principal root traits described above may be influenced by abiotic stresses during the rapid phase of root growth and expansion in the pre-flowering period with implications for resource capture during later seed filling, and these effects are now considered in the following section of this entry.

Effects of Abiotic Stress and Root Development

Drought and Root Development

Drought overall usually reduces the size of root systems. Water deficits decrease carbon assimilation by the



Roots and Uptake of Water and Nutrients. Figure 1
Interrelationships between these root traits and their relations with resource capture

plant due to a reduction of green leaf area, but also due to a decline in net photosynthetic rate. Nevertheless, under drought plants tend to increase the proportion of total carbon allocated to the roots [25, 45]. For example, plants responded to water deficits by increasing the proportion of assimilate allocated to roots in wheat [46, 47] and barley [48]. Experiments using pulse-labeled 13C in wheat have shown that water deficits increase the allocation of assimilated C to the roots due to a greater reduction of growth in the above-ground than below-ground plant components [49]. Although the relative root dry mass tends to increase with water deficits the absolute weight of both roots and shoots tends to decrease. As the soil dries, there are changes in its physical condition such as increases in soil strength [50]. Shoots are generally more affected by drought than roots, associated with more severe water deficits developing and persisting longer in the transpiring shoots [49, 51, 52]. Thus, roots are typically prioritized during drought to facilitate access to water while decreasing transpiration. The relationship between these two systems is often described as a competition where both roots and shoots compete for carbohydrates, minerals, and water, the most successful being the one nearer the source [24]. Therefore, the growth of the root and shoot systems is an integrative process working in a functional equilibrium [53, 54]. So when light is limited, root growth will be more restricted than shoots and the opposite happens when soil resources are in deficit; this functional balance hypothesis is elegantly explained by Brouwer [24]. In addition, under water-limiting conditions, solutes may accumulate in the root tip attracting the movement of water by diffusion, allowing the cells in the root tip to maintain their turgor and growth [55].

Although water deficits typically increase the percentage of carbon allocated to the roots, there are some reports of contrasting responses of root partitioning to drought amongst cultivars, for example, in glasshouse- [52] and field-grown wheat [56]. Therefore, the root growth response under drought is not simple, since drought not only affects plant and root growth but also the soil structure and N availability [57]. For example, water deficits may have a neutral effect on root weight, but still influence root length and its distribution with depth in wheat [47]; or increase SRL with drought in bread wheat [58] and durum wheat and barley [59]. There is some evidence that thinner roots may themselves be more vulnerable to drought [44, 60]. Therefore, the relatively high diameters reported for irrigated compared to droughted plants [58, 61] might relate to the necessity of the root system to support a larger plant and facilitate faster and greater water uptake and transport in well-watered conditions.

Leaf expansion and senescence are particularly susceptible to water deficiency [62]. The causes for restricted leaf expansion with drought have been discussed extensively, and there are mainly two views on the underlying mechanisms involved. Some authors attribute the cause to water relations (water potential and cell turgor) in the leaf [63, 64], while others attribute it to root chemical signals, such as abscisic acid (ABA), transmitted to the leaves, in response to water depletion in the soil [65–67]. ABA concentration increases in shoots, leaves, and roots in plants grown under water deficits and its exogenous application on well-watered plants mimics many of the drought effects on the plant [68, 69]. The chemical mechanism involves the synthesis of the plant hormone ABA by the roots when sensing the drying of the soil, and the transfer of ABA in the xylem to shoots and leaves inducing stomatal closure hence reducing water uptake and shoot and leaf growth [68, 70–72].

Evidence for water relations as the main cause of the decrease in leaf expansion was described by Boedt & Hensley (1987 in [63]) where leaves of field-grown maize showed visual symptoms of water stress in soil near field capacity. Tazaki et al. (1980 in [63]) in Japan reported similar effects for rice leaves, even though plants were rooted in wet soil. Furthermore, seedling experiments in maize plants using the pressure-pump technique [64] showed that an increase in the water pressure in the roots was quickly and fully transmitted to the base of the leaf increasing the leaf elongation. In contrast with these findings, Passioura [66] growing wheat seedlings in drying soil but maintaining leaf turgidity using the pressure-chamber method, showed a decrease in the relative expansion rate of leaves. Additional evidence for the root chemical signal was given by Gollan et al. [73] where wheat and sunflower plants showed a decrease in stomatal conductance with an increase of water deficits while the pressure in the plant was maintained. Using partial root-zone drying (PRD) techniques where half of the root system is droughted while the other half is irrigated, to maintain the same leaf water status as control plants (full irrigation), results showed a decrease of 65% of leaf area and 70% of water loss in apple plant seedlings subjected to PRD [67].

A more recent hypothesis is that both hydraulic and chemical signals interact and that the importance of one or the other will depend on the timescale considered [74]. Experiments in maize and barley showed that sudden changes in leaf water status by light, humidity, or salinity greatly affect leaf-elongation rate, and that those effects vanished when their roots were placed in a pressure chamber to maintain the xylem and air pressures in equilibrium, showing that hydraulic relations dominated in this response [75]. If the saline or water stress was prolonged, water relations were overridden by chemical signals and pressurization failed to maintain leaf elongation rates [75]. The combination of hydraulic and chemical factors was also demonstrated by differences in the sensitivity of different maize lines under drought to xylem ABA [76].

Nutrients and Root Development

It is well established that plants respond to N and P deficiencies by increasing RMR due to the functional equilibrium between the growth of the root and shoot [24, 77–81]. Crop root systems are plastic and respond by proliferating roots to exploit patches of nutrients where the distribution within the soil is uneven [82]. For example, responses to aqueous fertilizer in wheat have been observed within 24 h of application [83]. Frequently there is a strong association between root length and P uptake. Root proliferation in P-rich patches is, therefore, relatively straightforward to interpret in terms of a "foraging" response. The responses of roots to N- and P-rich patches of soil include proliferation of laterals and stimulation of nutrient inflow (uptake rate per unit root length) within the patch [81]. Nitrate uptake from a N-rich patch may compensate for an uneven supply of nitrate to the whole root system. Localized N application on barley seminal root systems promoted the number and extension rate of both first- and second-order lateral roots [84]. The potential magnitude of the responses to N and P has been demonstrated in barley by Drew and Saker [85, 86]. With regard to genetic effects, Zhang and Forde [87] demonstrated that, in Arabidopsis, the extension of lateral roots in nitrate-rich patches is partly under genetic control. Since irrigation and N fertilizer can cause root proliferation in the surface soil [47, 88], the distribution of the availability of these resources earlier in the crop's growth may alter the relative distribution of roots with depth (β) at anthesis.

For two barley varieties grown in Mediterranean field conditions, RMR increased under low N and P fertilizer supply compared to a control treatment with ample N and P supply [89]. Herrera et al. [90] in wheat showed that high N supply increased the number of roots, and when N was limited root formation ceased earlier. Barraclough et al. [47] observed an increase in RMR with low N supply in N x drought field experiments in winter wheat in the UK. N application effects on SRL are inconsistent, and increases, decreases, or neutral effects are reported for different species [44, 91]. Field experiments on spring barley and durum wheat in Jordan showed no consistent response for SRL for three different levels of N fertilizer [59]. SRL increased with N application under rain-fed conditions for durum wheat, but the opposite was found for spring barley. There appear to be few previous investigations regarding the effects of N fertilization on SRL and its components in cereals in field conditions. N application has been observed to increase mean root diameter in cereals but to decrease RW:RV [44, 91].

Root Traits and Resource Capture

The primary root traits for improved below-ground resource capture would appear to be root morphology (root axis number, rooting depth, rooting length density), root extension rates, root longevity, and root function along the length of the root system [9, 92, 93].

Water Capture

The importance of water for plants is unquestionable; it performs a varied number of physiological and structural functions. Water constitutes on average 80–90% of the fresh weight of herbaceous plants providing a continuous liquid phase in which gases, minerals, and other solutes enter the cells and move from one cell to another and within the different plant organs [94]. Water is a reactant or substrate in most of the plant's biochemical reactions (e.g., photosynthesis) and it maintains the plant turgor essential for cell growth, enlargement, form and movement of various plant structures, like the stomata opening [8, 94].

Crop production is closely related to water transpired; therefore, maintaining an uninterrupted supply of water to leaves is essential to maximize yields. Water capture is intimately related with root size, usually measured, as surface area, volume, or length. According to the theoretical model of van Noordwijk [95], the rate of water uptake by the plant is mainly limited by the transport in the soil toward the root (soil–root interface). Therefore, the density of roots, measured as length per unit soil volume (root length density – RLD, cm cm^{-3}), is the most suitable parameter to describe water uptake by plant roots. Prolific root systems are more effective at capturing water than sparse systems, but inter-root competition sets a natural ceiling on optimum RLD in cereals, above which further increases require excessive roots which do not have measurable effects on water uptake [95]. Theoretical calculations predict a critical root length density (C_{RLD}) of about 1 cm cm^{-3} for water uptake. This figure broadly concurs with the values reported for water uptake of Gregory and Brown [96] and Barraclough et al. [47] who showed that a RLD of 1 cm cm^{-3} was associated with the abstraction of all of the available water by both spring barley and winter wheat, respectively. However, for upland rice, values of C_{RLD} between 1.5 and 1.6 cm cm^{-3} have been reported [97, 98] and in controlled environment conditions values as low as 0.30 cm cm^{-3} [99].

RLD distribution with depth is principally determined by time for growth (residence times are greater in the topsoil than the subsoil), soil porosity and strength, and water availability [20]. Root length density in wheat is typically below the C_{RLD} of ca. 1 cm cm^{-3} at soil depths below ca. 80 cm [18, 21, 36, 100]. A modeling study concluded that distributing roots relatively deeper in the soil profile and decreasing SRL would confer greater water capture and yield under low water availability in wheat [30]. Experimental evidence also supports the strategy of distributing roots relatively deeper to improve water capture under drought. Synthetic derivative wheat lines showed increased water uptake associated with a root system that was distributed relatively deeper in the soil compared with recurrent parents [10], and the drought tolerance of spring wheat SeriM82 was related to its relatively deep root system compared to the check cultivar Hartog [101]. Further root traits which could be beneficial in boosting water capture include enhanced post-anthesis root longevity and root penetration ability [102], although there is relatively little information on genetic variation in these traits in cereals.

In rice under flooded conditions, attempts to optimize the root system through plant breeding methods must additionally allow for the complicated interplay between adaptations for internal aeration and those for efficient nutrient acquisition. A recent model developed by Kirk [103] provides a coherent representation of the rice root system in submerged soil and predicted that a system of coarse, aerenchmymatous primary roots with gas-impermeable walls conducting O_2 down to short, fine, gas-permeable laterals provided the best compromise between the need for internal aeration and the need for the largest possible absorbing surface per unit root mass.

Nutrient Capture

Water and nutrient uptake should really be considered together since nutrients become less available as the soil dries. Nitrate is readily leached down the soil profile and consequently rooting depth is an important attribute for soil N acquisition. For a long time, due to its high mobility in soil, N supply was considered independent of the root system characteristics, assuming that only mass flow and diffusion were the relevant mechanisms for the uptake of N by the plant [90]. The role of crop root systems in capturing N is still a topic of debate. Findings of Robinson et al. [104] indicated only 4–11% of the total root length is involved in N uptake. On the other hand, Palta and Watt [9] demonstrated through 15N labeling experiments on wheat that vigorous root systems captured ca. 60% more N in the top 0.2 m of the soil profile than non-vigorous root systems. Furthermore, a positive correlation was found between nitrate and water uptake and root length density in maize [105–107] and in several catch crop species [108]. These studies indicated that greater root length densities are generally more effective in N acquisition. However, root length is probably more important for the uptake of relatively immobile ions such as phosphates [109, 110]. However, some investigations have found uptake rates of phosphate, calcium, and potassium from solution are poorly related to root length [111, 112], possibly because root length is only significant if the uptake of these nutrients is limiting [7]. In addition, several investigations have shown that nitrate capture depends on the ability of the root system to respond to spatial and temporal nitrogen supply [81, 113].

The ability to capture N depends mainly on the amount of nitrate present in the soil relative to the morphology of the root system. Nitrate is supplied to the root system by mass flow (ions carried along in the transpiration stream) and diffusion (ions moving down a concentration gradient, either through bulk soil water or along water films surrounding particles). About 50% of the N taken up by wheat crops may be transported by mass flow [16]. As for water uptake, inter-root competition sets a natural ceiling on optimum RLD in cereals of about 1 cm cm^{-3} [95]. RLD distribution with depth is principally determined by time for growth (residence times are greater in the topsoil than the subsoil), soil porosity and strength, and nutrient and water availability [20]. The modeling study of King et al. [30] concluded that distributing roots relatively deeper in the soil profile and increasing SRL would confer greater N capture and yield under low N availability. The particular properties of each nutrient in the soil impose different RLD requirements for effective uptake. For example, due the low mobility of phosphorous (P) in the soil, a higher RLD of ca. 10 cm cm^{-3} is required for effective P uptake compared to water and/or nitrogen [114, 115]. Similar to water uptake, root traits that could be beneficial in boosting N capture include enhanced root longevity post-anthesis and root penetration ability [102], although there is relatively little information on genetic variation in these traits in wheat. Barraclough et al. [47] in N × drought field experiments in winter wheat in the UK found that water uptake increased with N due to a higher RLD and higher ground cover reducing soil evaporation. Positive correlations between nitrogen capture and RLD have also been found in maize [106, 107] and durum wheat and barley [59]. Although higher N supply tends to increases total RLD and N uptake by the crop, this generally results in a decrease in N uptake efficiency (crop N uptake/N available) [116, 117] leading to potentially greater losses of nitrate to the environment. A recent modeling exercise suggested that higher RLD and deeper rooting depths would reduce residual nitrate in high leaching soils [118]. Forde & Clarkson [119] concluded that there was no strong evidence for significant age-dependent changes in capacity of roots to absorb nitrate or ammonium ions.

Nutrient uptake may also be influenced by root membrane transporter systems. With regard to N uptake, recent work in *Arabidopsis* indicates nitrate is actively transported across the plasma membranes of plant cells, but net uptake is a balance between active influx and passive efflux. Two distinct gene families of nitrate transporters, *NRT1* and *NRT2*, have been identified [119–122] in the *Arabidopsis* genome. Some members of both *NRT1* and *NRT2* gene families are nitrate inducible, and are expressed in the root epidermis and in root hairs, and are likely to be responsible for the uptake of nitrate from the soil (e.g., [123–126]). There are prospects for transferring this information to wheat for improving efficiency of N uptake in the long term if the root screens used for *Arabidopsis* could be adapted to the larger and structurally different root system of wheat. An extensive review of this area is beyond the scope of this entry. Fortunately, excellent reviews are recently available in this topic area [127, 128].

With regard to the carbon costs of roots, it seems there is likely only a limited capacity to reduce root partitioning below current values of ca. 10% at anthesis in cereals in high yield potential ideotypes, due to the trade-off with water and N capture required for future biomass gains. However, a deeper relative distribution of roots while maintaining RMR could comprise part of an ideotype to maximize yield in future breeding programs.

Application of Rooting Traits in Breeding for Tolerance of Abiotic Stresses

Genetic variation in root system size has been widely reported in grain crops (e.g., [6, 45, 129]), but root distribution varies strongly with soil characteristics such as water and nutrient availability and mechanical impedance [54]. The RMR of wheat or barley is typically ca. 30% during early vegetative growth decreasing to ca. 10% by anthesis [18, 96, 130]. Effects of increasing plant height on root partitioning have been studied using isolines and are generally either neutral or negative in wheat [22, 130–132].

Breeding cultivars better adapted to particular conditions of drought and/or low nutrient availability will play a major role for the future of crop production for adaptation to climate change. Breeding for more effective use of water or nutrients while maintaining, or ideally, increasing yield potential is a difficult task. It will be important in the most efficient crop systems to combine optimized agronomy and new cultivars efficient at acquiring below-ground resources. To date there are relatively few examples of root morphological or anatomical traits that have been successfully selected for crop breeding programs to result in improved performance. In a recent review on this topic, Palta and Watt [9] cited five examples of rooting traits directly related to below-ground resource capture in breeding, including long root hairs for increased P uptake in barley [12], reduced xylem vessel diameter in seminal roots of wheat [13], and increased root density at depth through faster root extension rate in wheat [133]. In addition, improved resource capture has been achieved by alleviating other stresses on roots, for example, resistance to cereal cyst nematode in wheat [134]. Another encouraging example of introgression of rooting traits in crop breeding can be found in rice. When near isogenic lines (NILs) obtained through marker-assisted backcrossing for four QTLs for root length were field-tested, they outperformed the recurrent parent for yield and biomass [135]. In order to introgress rooting traits into elite genotypes, it will be necessary to identify genetic diversity for the key traits as well as developing methods for rapid high-throughput screening of lines in breeders trials.

Field phenotyping methods for roots in cereals were reviewed by Manske et al. [136] and Polomski and Kuhn [137], including the use of rhizotrons and assessments of root parameters from soil cores (root washing and root counts/image analysis). Although several screening tests have been designed to generate accurate and robust data from seedling plants grown under artificial conditions, these phenotypes can rarely be extrapolated to field conditions because of the pronounced plasticity of root growth and development processes [138]. Field phenotyping for root traits in breeding programs is currently infeasible, so genetic progress will depend on the development of high-throughput controlled-environment screens or molecular markers for root traits for marker-assisted selection (MAS). The use of root-observation chambers and a nondestructive digital imaging technique offers some promise [139], but may be less suitable for screening of root traits that are expressed at later stages of crop development.

Future Directions

The existence of significant genetic variation for rooting traits has not resulted to date, except for a few exceptions as mentioned above, in the incorporation of rooting traits in conventional breeding. Nevertheless, future genetic progress for resistance to abiotic stresses should be accelerated by the fact that the genetic control of rooting traits can now be revealed through the application of rapidly emerging genetic resources facilitating the fine mapping of root QTL. Indeed, the cloning of the first root QTL is ongoing. However, successful exploitation of genomics tools and strategies in plant breeding requires extensive and precise phenotyping of agronomic traits for breeding materials and mapping populations. The capacity for precise phenotyping under reliable conditions probably represents the most limiting factor for the progress of genomic studies on root traits underlying resilience to abiotic stresses. There is a need for a high precision because the differences may be small, and detailed physiological measurements (e.g., of growth rate) are difficult when a large numbers of genotypes are involved.

Physiological perspectives that require more attention are analyses to measure the full carbon costs of the turnover of root material (and therefore of the root system), which are presently poorly quantified especially in environments where soil stresses are common. The role of organic substances in the rhizosphere secreted by the root that are able to modify the environment to secure improved water uptake requires more attention. Future research should focus on the importance of root plasticity for nutrient capture rather than simply measuring the size of the response. More studies at the plant community level rather than on single plants are required to translate fundamental studies on root growth and function to the improved water and nutrient capture at the crop scale.

Additionally, future work should aim to address the potential use of marker-assisted backcrossing for root QTLs and to exploit findings in *Arabidopsis* where root screens for mutants have identified genes such as AUX1 and LAX3 that regulate important root architectural traits such as lateral root development [140]. There is a continuing need to integrate "omics" technologies with plant physiology, agronomy, breeding, and disciplines related to the rhizosphere. In the future to meet the challenge of raising biomass and yield potential as well as improving resilience to abiotic stresses it will be crucial that new root research fosters collaborations between breeders, geneticists, physiologists, crop physiologists, and soil scientists (among others) to translate the genetic data generated from the new genomics resources into improved crop performance.

Bibliography

Primary Literature

1. Lutz W, Sanderson W, Scherbov S (2001) The end of world population growth. Nature 412(6846):543–545
2. Hamilton R (2009) Agriculture's sustainable future: breeding better Crops. In: Scientific American: http://www.scientificamerican.com/article.cfm?id=agricultures-sustainable-future
3. Hirel B, Le Gouis J, Ney B, Gallais A (2007) The challenge of improving nitrogen use efficiency in crop plants: towards a more central role for genetic variability and quantitative genetics within integrated approaches. J Exp Bot 58(9):2369–2387
4. Cattivelli L, Rizza F, Badeck FW, Mazzucotelli E, Mastrangelo AM, Francia E, Marë C, Tondelli A, Stanca AM (2008) Drought tolerance improvement in crop plants: an integrated view from breeding to genomics. Field Crops Res 105(1–2):1–14
5. Raun WR, Solie JB, Johnson GV, Stone ML, Mullen RW, Freeman KW, Thomason WE, Lukina EV (2002) Improving nitrogen use efficiency in cereal grain production with optical sensing and variable rate application. Agron J 94:815–820
6. O'toole JC, Bland WL (1987) Genotypic variation in crop plant-root systems. Adv Agron 41:91–145
7. Hoad SP, Russell G, Lucas ME, Bingham IJ (2001) The management of wheat, barley and oat root systems. Adv Agron 74:193–246
8. Gregory PJ (2006) Plant roots: growth, activity, and interaction with soils. Blackwell, Oxford

9. Palta J, Watt M (2009) Vigorous crop root systems: form and function for improving the capture of water and nutrients. In: Crop Physiology – Applications for genetic improvement and agronomy, Elsevier, San Diego
10. Reynolds M, Dreccer F, Trethowan R (2007) Drought-adaptive traits derived from wheat wild relatives and landraces. J Exp Bot 58(2):177–186
11. Yoshida S, Hasegawa S (1982) The rice root system its development and function. In: O'Toole JC (ed) Drought resistance in crops with emphasis on rice. IRRI, Manila
12. Gahoonia TS, Nielsen NE (2004) Barley genotypes with long root hairs sustain high grain yields in low-P field. Plant Soil 262(1):55–62
13. Richards RA, Passioura JB (1989) A breeding program to reduce the diameter of the major xylem vessel in the seminal roots of wheat and its effect on grain yield in rain-fed environments. Aust J Agric Res 40(5):943–950
14. Fitter AH (1985) Functional significance of root morphology and root system architecture. In: Fitter AH, Atkinson D, Read DJ, Usher MB (eds) Ecological interactions in soil, plants, microbes and animals. Blackwell, Oxford, pp 87–106
15. Lynch J (1995) Root architecture and plant productivity. Plant Physiol 109:7–13
16. Gregory PJ, Crawford DV, Mcgowan M (1979) Nutrient relations of winter wheat. 2. Movement of nutrients to the root and their uptake. J Agri Sci 93:495–504
17. Key JM (1973) The wheat root. In: Sears ER, Sears LMS (eds) 4th International wheat genetics symposium pp 827–842. Agricultural Experiment Station, College of Agriculture, University of Missouri, Columbia, Missouri, Columbia
18. Gregory PJ, Mcgowan M, Biscoe PV, Hunter B (1978) Water relations of winter wheat.1. Growth of root system. J Agric Sci 91:91–102
19. Klepper B, Belford RK, Rickman RW (1984) Root and shoot development in winter wheat. Agron J 76(1):117–122
20. Barraclough PB, Weir AH, Kuhlmann H (1991) Factors affecting the growth and distribution of winter wheat roots under UK field conditions. Dev Agr Manag For Ecol 24:410–441
21. Barraclough PB, Weir AH (1988) Effects of compacted subsoil layer on root and shoot growth, water use and nutrient uptake of winter wheat. J Agric Sci 110:207–216
22. Siddique KHM, Belford RK, Tennant D (1990) Root:shoot ratios of old and modern, tall and semidwarf wheats in a mediterranean environment. Plant Soil 121(1):89–98
23. Barraclough PB (1984) The growth and activity of winter-wheat roots in the field – root-growth of high-yielding crops in relation to shoot growth. J Agri Sci 103:439–442
24. Brouwer R (1983) Functional equilibrium: sense or nonsense? Neth J Agric Sci 31(4):335–348
25. Gregory PJ, Palta JA, Batts GR (1997) Root systems and root: mass ratio – carbon allocation under current and projected atmospheric conditions in arable crops. Plant Soil 187(2): 221–228
26. Gerwitz A, Page ER (1974) An empirical mathematical model to describe plant root systems. J Appl Ecol 11(2):773–781
27. Gale MR, Grigal DF (1987) Vertical root distributions of northern tree species in relation to successional status. Can J For Res 17(8):829–834
28. Robertson MJ, Fukai S, Hammer GL, Ludlow MM (1993) Modelling root growth of grain sorghum using the CERES approach. Field Crops Res 33:113–130
29. Zhuang J, Yu GR, Nakayama K (2001) Scaling of root length density of maize in the field profile. Plant Soil 235:135–142
30. King J, Gay A, Sylvester-Bradley R, Bingham I, Foulkes J, Gregory P, Robinson D (2003) Modelling ceral root systems for water and nitrogen capture: towards an economic optimum. Ann Bot 91:383–390
31. Thomas Fukai S, Hammer GL (1995) Growth and yield response of barley and chickpea to water stress under three environments in Southeast Queensland. II* Root growth and soil wate extraction pattern. Aust J Agric Res 46:35–48
32. Monteith JL (1986) How do crops manipulate water-supply and demand. Philos Trans R Soc Lond Ser Math Phys Eng Sci 316(1537):245–259
33. Robertson MJ, Fukai S, Ludlow MM, Hammer GL (1993) Water extraction by grain sorghum in a sub-humid environment. II. Extraction relation root growth. Field Crops Res 33:99–112
34. Eissenstat DM (1991) On the relationship between specific root length and the rate of root proliferation - a field-study using citrus rootstocks. New Phytol 118(1):63–68
35. Welbank PJ, Gibb MJ, Taylor PJ, Williams ED (1974) Root growth of cereal crops. In: Rothamsted Experimental Station Report, 1973, Part 2 pp 26–66
36. Barraclough PB, Leigh RA (1984) The growth and activity of winter wheat roots in the field: the effect of sowing date and soil type on root growth of high yielding crops. J Agric Sci 103:59–74
37. Van Noordwijk M, Brouwer G (1991) Review of quantitative root length data in agriculture. Elsevier, Amsterdam, Pays-Bas
38. Ryser P (1996) The importance of tissue density for growth and life span of leaves and roots: A comparison of five ecologically contrasting grasses. Funct Ecol 10(6):717–723
39. Wahl S, Ryser P (2000) Root tissue structure is linked to ecological strategies of grasses. New Phytol 148:459–471
40. Valenzuela-Estrada LR, Vera-Caraballo V, Ruth LE, Eissenstat DM (2008) Root anatomy, morphology, and longevity among root orders in *Vaccinium corymbosum* (*Ericaceae*). Am J Bot 95(12):1506–1514
41. Eissenstat DM (1992) Costs and benefits of constructing roots of small diameter. J Plant Nutr 15(6–7):763–782
42. Fitter AH (1996) Characteristics and functions of root systems. In: Waisel Y, Eshel A, Kafkafi U (eds) Plant roots: the hidden half. Marcel Dekker, New York, pp 1–20
43. Hetrick BAD, Leslie JF, Wilson GT, Kitt DG (1988) Physical and topological assessment of effects of a vesicular-arbuscular mycorrhizal fungus on root architecture of big bluestem. New Phytol 110(1):85–96
44. Ryser P (1998) Intra- and interspecific variation in root length, root turnover and the underlying parameters. In: Lambers H, Poorter H, van Vuuren MMI (eds) Inherent variation in plant

growth. Physiological mechanisms and ecological consequences. Backhuys, Leiden

45. Ehdaie B, Waines JG (2003) 1RS translocation increases root biomass in Veery-type wheat isogenic lines and associates with grain yield. In: Pogna NE, Romano M, Pogna EA, Galterio G (eds) Proceedings of the 10th International wheat genetics symposium, ISC Paestum, Italy, Rome, pp 693–695
46. Hamblin A, Tennant D, Perry MW (1990) The cost of stress - dry-matter partitioning changes with seasonal supply of water and nitrogen to dryland wheat. Plant Soil 122(1):47–58
47. Barraclough PB, Kuhlmann H, Weir AH (1989) The effects of prolonged drought and nitrogen-fertilizer on root and shoot growth and water-uptake by winter-wheat. J Agron Crop Sci 163(5):352–360
48. Zenisceva L (1990) The importance of the root-system in adaptation of spring barley genotypes to the conditions of environment. Rost Vyroba 36:937–945
49. Palta JA, Gregory PJ (1997) Drought affects the fluxes of carbon to roots and soil in C-13 pulse-labelled plants of wheat. Soil Biol Biochem 29(9–10):1395–1403
50. Martino DL, Shaykewich CF (1994) Root penetration profiles of wheat and barley as affected by soil penetration resistance in field conditions. Can J Soil Sci 74:193–200
51. Kramer PJ (1983) Water relations of plants. Academic, London
52. Karrou M, Maranville JW (1994) Response of wheat cultivars to different soil nitrogen and moisture regimes: I. Dry matter partitioning root growth. J Plant Nutr 17(5):729–744
53. Davidson RL (1969) Effect of root/leaf temperature differentials on root/shoot ratios in some pasture grasses and clover. Ann Bot 33(3):561–569
54. Gregory PJ (1994) Root growth and activity. In: Boote KJ, Bennett JM, Sinclair TR, Paulsen GM (eds) Physiology and determination of crop yield. Soil Science Soc America, Madison, pp 65–93
55. Sharp RE, Davies WJ (1979) Solute regulation and growth by roots and shoots of water-stressed maize plants. Planta 147(1):43–49
56. Li FM, Yan X, Li FR, Guo AH (2001) Effects of different water supply regimes on water use and yield performance of spring wheat in a simulated semi-arid environment. Agric Water Manage 47(1):25–35
57. Robinson D (1994) Resource capture by single roots. In: Monteith JL, Scott RK, Unsworth MH (eds) Resource capture by crops. Nottingham University Press, Nottingham, pp 53–76
58. Baburai Nagesh AK (2006) The physiological and genetic bases of water-use efficiency in winter wheat. Ph.D. thesis, University of Nottingham
59. Ebrahim NM (2008) Responses of root and shoot growth of durum wheat (Triticum turgidum L. var durum) and barley (Hordeum vulgare L.) plants to different water and nitrogen levels. PhD Thesis University of Jordan, Jordan, 284 pp
60. Fitter AH (1987) An architectural approach to the comparative ecology of plant root systems. New Phytol 106:61–77
61. Li FM, Liu XL, Li SQ (2001) Effects of early soil water distribution on the dry matter partition between roots and shoots of winter wheat. Agric Water Manage 49(3):163–171
62. Turner NC, Begg JE (1981) Plant-water relations and adaptation to stress. Plant Soil 58(1–3):97–131
63. Kramer PJ (1988) Changing concepts regarding plant water relations. Plant Cell Environ 11(7):565–568
64. Hsiao TC, Frensch J, Rojas-Lara BAR (1998) The pressure-jump technique shows maize leaf growth to be enhanced by increases in turgor only when water status is not too high. Plant Cell Environ 21(1):33–42
65. Passioura JB (1988) Changing concepts regarding plant water relations - response. Plant Cell Environ 11(7):569–571
66. Passioura JB (1988) Root signals control leaf expansion in wheat seedlings growing in drying soil. Aust J Plant Physiol 15(5):687–693
67. Gowing DJG, Davies WJ, Jones HG (1990) A positive root-sourced signal as an indicator of soil drying in Apple, *Malus x domestica* Borkh. J Exp Bot 41(233):1535–1540
68. Davies WJ, Tardieu F, Trejo CL (1994) How do chemical signals work in plants that grow in drying soil? Plant Physiol 104(2):309–314
69. Taylor IB, Burbidge A, Thompson AJ (2000) Control of abscisic acid synthesis. J Exp Bot 51(350):1563–1574
70. Zhang J, Davies WJ (1990) Changes in the concentration of aba in xylem sap as a function of changing soil-water status can account for changes in leaf conductance and growth. Plant Cell Environ 13(3):277–285
71. Davies WJ, Zhang J (1991) Root signals and the regulation of growth and development of plants in drying soil. Annu Rev Plant Physiol Plant Mol Biol 42(1):55–76
72. Hartung W, Jeschke WD (1999) Abscisic acid: a long-distance stress signal in salt-stressed plants. In: Lerner HR (ed) Plant responses to environmental stresses. Marcel Dekker, New York, pp 333–348
73. Gollan T, Passioura JB, Munns R (1986) Soil water status affects the stomatal conductance of fully turgid wheat and sunflower leaves. Aust J Plant Physiol 13(4):459–464
74. Aphale SL (2004) Role of root to shoot signalling in coordinating responses to nitrogen deficiency. Ph.D. thesis, University of Nottingham
75. Munns R, Passioura JB, Guo JM, Chazen O, Ramer GR (2000) Water relations and leaf expansion: importance of time scale. J Exp Bot 51(350):1495–1504
76. Stikić R, Davies WJ (2000) Stomatal reactions of two different maize lines to osmotically induced drought stress. Biol Plant 43(3):399–405
77. Novoa R, Loomis RS (1981) Nitrogen and plant production. Plant Soil 58(1–3):177–204
78. Dreccer M, Schapendonk A, Slafer G, Rabbinge R (2000) Comparative response of wheat and oilseed rape to nitrogen supply: absorption and utilisation efficiency of radiation and nitrogen during the reproductive stages determining yield. Plant Soil 220(1):189–205

79. Mengel K, Kirkby EA (2001) Principles of plant nutrition. Kluwer, London
80. Reich PB (2002) Root-shoot relations: optimality in acclimation and adaptation or the: "Emperor's new clothes"? In: Waisel Y, Eshel A, Kafkafi U (eds) Plant roots: the hidden half. Marcel Dekker, New York, pp 205–220
81. Robinson D (2001) Root proliferation, nitrate inflow and their carbon costs during nitrogen capture by competing plants in patchy soil. Plant Soil 232(1–2):41–50
82. Robinson D, Hodge A, Griffiths BS, Fitter AH (1999) Plant root proliferation in nitrogen-rich patches confers competitive advantage. Proc R Soc Lond B Biol Sci 266(1418):431–435
83. Jackson RB, Caldwell MM (1989) The timing and degree of root proliferation in fertile-soil microsites for 3 cold-desert perennials. Oecologia 81(2):149–153
84. Drew MC, Saker LR, Ashley TW (1973) Nutrient supply and the growth of the seminal root system in barley: I. Effect nitrate concentration growth axes laterals. J Exp Bot 24(6): 1189–1202
85. Drew MC, Saker LR (1975) Nutrient supply and the growth of the seminal root system in barley. II. Localized, compensatory increases in lateral root growth and rates op nitrate uptake when nitrate supply is restricted to only part of the root system. J Exp Bot 26(1):79–90
86. Drew MC, Saker LR (1978) Nutrient supply and the growth of the seminal root system in barley. III. Compensatory increases in growth of lateral roots, and in rates of phosphate uptake, in response to a localized supply of phosphate. J Exp Bot 29(2):435–451
87. Zhang HM, Forde BG (1998) An Arabidopsis MADS box gene that controls nutrient-induced changes in root architecture. Science 279(5349):407–409
88. Robinson D (1994) The responses of plants to non-uniform supplies of nutrients. New Phytol 127(4):635–674
89. Brown SC, Keatinge JDH, Gregory PJ, Cooper PJM (1987) Effects of fertilizer, variety and location on barley production under rainfed conditions in northern syria 1root shoot growth. Field Crops Res 16:53–66
90. Herrera JM, Stamp P, Liedgens M (2005) Dynamics of root development of spring wheat genotypes varying in nitrogen use efficiency. In: Foulkes J, Russel G, Hawkesford M, Gooding M, Sparkes D, Stockdale E (eds) Roots and the soil environment II. Association of Applied Biologists, Warwick, pp 197–201
91. Ryser P, Lambers H (1995) Root and leaf attributes accounting for the performance of fast- and slow-growing grasses at different nutrient supply. Plant Soil 170:251–265
92. Foulkes MJ, Hawkesford MJ, Barraclough PB, Holdsworth MJ, Kerr S, Kightley S, Shewry PR (2009) Identifying traits to improve the nitrogen economy of wheat: recent advances and future prospects. Field Crops Res 114:329–342
93. Doussan C, Pages L, Pierret A (2003) Soil exploration and resource acquisition by plant roots: an architectural and modelling point of view. Agronomie 23(5–6):419–431
94. Kramer PJ, Boyer JS (1995) Water relations of plants and soils. Academic, San Diego
95. Van Noordwijk M (1983) Functional interpretation of root densities in the field for nutrient and water uptake. In: Böhm W, Kutschera L, Lichtenegger E (eds) Root ecology and its practical application, Interantinal Symposium. Gumpenstein 1982, Irdning, Bundesanstalt Gumpenstein, pp 207–226
96. Gregory PJ, Brown SC (1989) Root growth, water use and yield of crops in dry environments: what characteristics are desirable? Aspects Appl Biol 22:235–243
97. Pantuwan G, Fukai S, Cooper M, O'toole JC, Sarkarung S (1997) Root traits to increase drought resistance in rainfed lowland rice. In: Fukai S, Cooper M, Salisbury J (eds) Proceedings of Breeding Strategies for Rainfed Lowland Rice in Drought-Prone Environments, vol 77, Canberra, pp 170–179
98. Lilley JM, Fukai S (1994) Effect of timing and severity of water deficit on four diverse rice cultivars I. rooting pattern soil water extraction. Field Crops Res 37(3):205–213
99. Siopongco J, Yamauchi A, Salekdeh H, Bennett J, Wade LJ (2005) Root growth and water extraction response of doubled-haploid rice lines to drought and rewatering during the vegetative stage. Plant Prod Sci 8(5):497–508
100. Ford KE, Gregory PJ, Gooding MJ, Pepler S (2006) Genotype and fungicide effects on late-season root growth of winter wheat. Plant Soil 284(1–2):33–44
101. Christopher JT, Manschadi AM, Hammer GL, Borrell AK (2008) Developmental and physiological traits associated with high yield and stay-green phenotype in wheat. Aust J Agric Res 59(4):354–364
102. Bengough AG, Bransby MF, Hans J, Mckenna SJ, Roberts TJ, Valentine TA (2006) Root responses to soil physical conditions; growth dynamics from field to cell. J Exp Bot 57(2): 437–447
103. Kirk GJD (2003) Rice root properties for internal aeration and efficient nutrient acquisition in submerged soil. New Phytol 159(1):185–194
104. Robinson D, Linehan DJ, Caul S (1991) What limits nitrate uptake from soil. Plant Cell Environ 14(1):77–85
105. Cooper PJM, Gregory PJ, Keatinge JDH, Brown SC (1987) Effects of fertilizer, variety and location on barley production under rainfed conditions in Northern Syria. 2: Soil water dyamics and crop water use. Field Crops Res 16:67–84
106. Wiesler F, Horst WJ (1993) Differences among maize cultivars in the utilization of soil nitrate and the related losses of nitrate through leaching. Plant Soil 151:193–203
107. Wiesler F, Horst WJ (1994) Root growth and nitrate utilization of maize cultivars under field conditions. Plant Soil 163: 267–277
108. Thorup-Kristensen K (1993) Root development of nitrogen catch crops and of a succeeding crop of broccoli. Acta Agriculturae Scaninavica B Soil Plant Sci 43:58–64

109. Gregory PJ (1994) Resource capture by root networks. In: Monteith JL, Scoot RK, Unsworth MH (eds) Resource capture by crops. Nottingham University Press, Nottingham, pp 77–97
110. Nielsen NE, Schjorring JK (1983) Efficiency and kinetics of phosphorus uptake from soil by various barley genotypes. Plant Soil 72(2–3):225–230
111. Gregory PJ (1994) Resource capture by root networks. In: Monteith JL, Scott RK, Unsworth MH (eds) Resource capture by crops. Nottingham University Press, pp 77–97
112. Gao S, Pan WL, Koenig RT (1998) Wheat root growth responses to enhanced ammonium supply. Soil Sci Soc America J 62:1736–1740
113. Fitter AH, Stickland TR (1991) Architectural analysis of plant-root systems. 2. Influence of nutrient supply on architecture in contrasting plant-species. New Phytol 118(3):383–389
114. López-Bucio J, Hernández-Abreu E, Sánchez-Calderón L, Nieto-Jacobo MF, Simpson J, Herrera-Estrella L (2002) Phosphate availability alters architecture and causes changes in hormone sensitivity in the arabidopsis root system. Plant Physiol 129(1):244–256
115. Poirier Y, Bucher M (2002) Phosphate transport and homeostasis in Arabidopsis. American Society of Plant Biologists, Rockville
116. Cabrera-Bosquet L, Molero G, Bort J, Nogués S, Araus JL (2007) The combined effect of constant water deficit and nitrogen supply on WUE, NUE and D13C in durum wheat potted plants. Ann Appl Biol 151(3):277–289
117. Pask A (2009) Optimising nitrogen storage in wheat canopies for genetic reduction in fertiliser nitrogen inputs. Ph.D. thesis, University of Nottingham
118. Dunbabin V, Diggle A, Rengel Z (2003) Is there an optimal root architecture for nitrate capture in leaching environments? Plant Cell Environ 26:835–844
119. Forde BG, Clarkson DT (1999) Nitrate and ammonium nutrition of plants: physiological and molecular perspectives. In: Ellow JA (ed) Advances in botanical research incorporating advances in plant pathology, vol 30. Academic, San Diego, pp 1–90
120. Crawford NM, Glass ADM (1998) Molecular and physiological aspects of nitrate uptake in plants. Trends Plant Sci 3(10):389–395
121. Williams LE, Miller AJ (2001) Transporters responsible for the uptake and partitioning of nitrogenous solutes. Annu Rev Plant Physiol Plant Mol Biol 52:659–688
122. Hawkesford MJ, Miller AJ (2004) Ion-coupled transport of inorganic solutes. In: Blatt MR (ed) Membrane transport in plants annual reviews 15. Blackwell/CRC Press, Oxford, pp 105–134
123. Lauter FR, Ninnemann O, Bucher M, Riesmeier JW, Frommer WB (1996) Preferential expression of an ammonium transporter and of two putative nitrate transporters in root hairs of tomato. Proc Natl Acad Sci USA 93(15):8139–8144
124. Zhuo DG, Okamoto M, Vidmar JJ, Glass ADM (1999) Regulation of a putative high-affinity nitrate transporter (Nrt2;1At) in roots of Arabidopsis thaliana. Plant J 17(5): 563–568
125. Ono F, Frommer WB, Von Wiren N (2000) Coordinated diurnal regulation of low- and high-affinity nitrate transporters in tomato. Plant Biol 2(1):17–23
126. Orsel M, Krapp A, Daniel-Vedele F (2002) Analysis of the NRT2 nitrate transporter family in Arabidopsis. structuer and gene expression. Plant Physiol 129(2):886–896
127. Bucher M (2007) Functional biology of plant phosphate uptake at root and mycorrhiza interfaces. New Phytol 173(1):11–26
128. Miller AJ, Shen Q, Xu G (2009) Freeways in the plant: transporters for N, P and S and their regulation. Curr Opin Plant Biol 12(3):284–290
129. Waines JG, Ehdaie B (2007) Domestication and crop physiology: roots of green-revolution wheat. Ann Bot 100(5): 991–998
130. Miralles DJ, Slafer GA, Lynch V (1997) Rooting patterns in near-isogenic lines of spring wheat for dwarfism. Plant Soil 197(1):79–86
131. Bush MG, Evans LT (1988) Growth and development in tall and dwarf isogenic lines of spring wheat. Field Crops Res 18(4):243–270
132. Mccaig TN, Morgan JA (1993) Root and shoot dry-matter partitioning in near-isogenic wheat lines differing in height. Can J Plant Sci 73(3):679–689
133. Hurd EA (1974) Phenotype and drought tolerance in wheat. Agric Meteorol 14(1–2):39–55
134. Ogbonnaya FC, Seah S, Delibes A, Jahier J, Lopez-Brana I, Eastwood RF, Lagudah ES (2001) Molecular-genetic characterisation of a new nematode resistance gene in wheat. Theor Appl Genet 102(4):623–629
135. Steele KA, Price AH, Shashidhar HE, Witcombe JR (2006) Marker-assisted selection to introgress rice QTLs controlling root traits into an Indian upland rice variety. Theor Appl Genet 112(2):208–221
136. Manske GGB, Ortiz-Monasterio JI, Vlek PLD (2001) Thecniques for measuring genetic diversity in roots. In: Reynods MP, Ortiz-Monasterio JI, McNab A (eds) Application of physiology in wheat breeding. Mexico, Cimmiyt
137. Polomski J, Kuhn N (2002) Root research methods. In: Waisel Y, Eshel A, Kafkafi U (eds) Plant roots: the hidden half. Marcel Dekker, New York, pp 295–321
138. De Dorlodot S, Forster B, Pages L, Price A, Tuberosa R, Draye X (2007) Root system architecture: opportunities and constraints for genetic improvement of crops. Trends Plant Sci 12:474–481
139. Manschadi A, Hammer G, Christopher J, Devoil P (2008) Genotypic variation in seedling root architectural traits and implications for drought adaptation in wheat (Triticum aestivum L). Plant Soil 303(1):115–129

140. Swarup K, Benkova E, Swarup R, Casimiro I, Peret B, Yang Y, Parry G, Nielsen E, De Smet I, Vanneste S, Levesque MP, Carrier D, James N, Calvo V, Ljung K, Kramer E, Roberts R, Graham N, Marillonnet S, Patel K, Jones JDG, Taylor CG, Schachtman DP, May S, Sandberg G, Benfey P, Friml J, Kerr I, Beeckman T, Laplaze L, Bennett MJ (2008) The auxin influx carrier LAX3 promotes lateral root emergence. Nat Cell Biol 10(8):946–954

Books and Reviews

Bassirirad H (2000) Kinetics of nutrient uptake by roots: responses to global change. New Phytol 147(1):155–169

Beeckman T (ed) (2010) Root development. Blackwell, Chichester

Bingham IJ (2001) Soil-root-canopy interactions. Ann Appl Biol 138(2):243–251

Bingham IJ, Hoad SP (2000) Towards below ground management In: HGCA conference: crop management into the millennium, Homerton College, Conference Centre, HGCA, Cambridge, pp 7.1–7.8

Bingham IJ, Foulkes MJ, Gay AP, Gregory PJ, King JA, Robinson D, Bradley RS (2002) Balancing root and canopy growth. In: HGCA (ed) R & D conference: agronomic intelligence: the basis for profitable production, pp 6.1–6.14

Blum A (ed) (2009) Plant stress http://plantstress.com/

Dakora FD, Phillips DA (2002) Root exudates as mediators of mineral acquisition in low-nutrient environments. Plant Soil 245(1):35–47

Lynch JP (2007) Roots of the second green revolution. Aust J Bot 55(5):493–512

Lynch J, Brown KM (eds) (2009) Root research methods. http://roots.psu.edu/en/methods

Schachtman DP, Goodger JQD (2008) Chemical root to shoot signaling under drought. Trends Plant Sci 13(6):281–287

Smit AL, Bengough AG, Engels C, Noordwijk MV, Pllerin S, Geijn SCVD (eds) (2000) Root methods. A handbook. Springer, Berlin

Van Noordwijk M, Martikainen P, Bottner P, Cuevas E, Rouland C, Dhillion SS (1998) Global change and root function. Glob change Biol 4:759–772

Waisel Y, Eshel A, Kafkafi U (eds) (2002) Plant roots: the hidden half. Marcel Dekker, New York

Wang E, Smith CJ (2004) Modelling the growth and water uptake function of plant root systems: a review. Aust J Agric Res 55:501–523